ROZUMNOSŤ

od algoritmov po zombie

Eliezer Yudkowsky

Rozumnosť

od algoritmov po zombie

Eliezer Yudkowsky

vydal v roku 2015
Machine Intelligence Research Institute
[Výskumný ústav strojovej inteligencie]
Berkeley 94704
Spojené štáty americké
intelligence.org

Eliezer Yudkowsky je vedecký pracovník v Machine Intelligence Research Institute.

preložil Viliam Búr
viliam@bur.sk

Machine Intelligence Research Institute ďakuje za štedrú pomoc všetkým, ktorí sa zúčastnili na uverejnení tejto knihy, a darcom, ktorí podporujú prácu MIRI vo všeobecnosti.

Prekladateľ ďakuje všetkým, ktorí prispeli spätnou väzbou k slovenskému prekladu.

Obsah

Kniha II. – Ako naozaj zmeniť svoj názor

L: Jednoduchá matematika evolúcie

M: Krehké ciele

N: Návod k ľudským slovám

Kniha IV. – Obyčajná skutočnosť

Kniha V. – Obyčajné dobro

Predslov

Vo svojich rukách držíte súhrn dvoch rokov každodenných článkov na blogu. Dodatočne sa obzerám späť na tento projekt a vidím mnoho vecí, ktoré som urobil celkom zle. Som s tým spokojný. Keby som sa obzrel späť a *nevidel* hromadu vecí, ktoré som urobil zle, znamenalo by to, že sa ani moje písanie ani moje rozmýšľanie od roku 2009 nezlepšilo. *Joj* je ten zvuk, ktorý vydávame, keď zlepšíme svoje názory a stratégie; takže obzrieť sa do minulosti a nevidieť na tom, čo ste urobili, nič nesprávne, znamená, že ste sa odvtedy nič nenaučili ani nezmenili svoj názor.

Bola chyba, že som svoje dva roky článkov na blogu nenapísal s úmyslom pomáhať ľuďom, aby sa im darilo lepšie v ich každodennom živote. Napísal som ich s úmyslom pomáhať ľuďom riešiť veľké, ťažké, dôležité problémy, a ako príklady som si vyberal dôležito vyzerajúce, abstraktné problémy.

Keď sa obzerám, toto bola druhá najväčšia chyba v mojom prístupe. Súvisí s *najväčšou* chybou v mojom písaní, totiž, že som si neuvedomil, že veľkým problémom pri učení sa tohto hodnotného spôsobu myslenia je zistiť, ako na to v praxi, nie poznať teóriu. Neuvedomil som si, že táto časť je priorita; na to dnes môžem povedať iba „Joj“ a „No jasné“.

Áno, niekedy sú tieto veľké veci naozaj veľké a naozaj dôležité; ale to nemení základnú pravdu, že zručnosti sa učia precvičovaním, a že je ťažšie precvičovať veci, ktoré sú vzdialené. (Dnes Center for Applied Rationality pracuje na napravení tejto mojej chyby systematickejším spôsobom.)

Tretia veľká chyba, ktorú som urobil, bolo príliš sa zameriavať na rozumné myslenie, a príliš málo na rozumné konanie.

Štvrtá najväčšia chyba, ktorú som urobil, bolo, že som mal lepšie zorganizovať obsah, ktorý som predkladal v postupnostiach. Konkrétne som mal omnoho skôr vytvoriť wiki a zjednodušiť tým čítanie článkov v postupnosti.

Aspoň *táto* chyba sa dá napraviť. V tomto diele Rob Bensinger zmenil poradie článkov a usporiadal ich, nakoľko len mohol bez snahy prepísať samotný materiál (hoci aj z toho časť prepísal).

Moja piata veľká chyba bola, že som sa – zo svojho pohľadu – snažil hovoriť otvorene o hlúposti toho, čo mi pripadalo ako hlúpe myšlienky. Snažil som sa vyhnúť klamu zvanému bulverizmus, kde svoju diskusiu *začnete* rozprávaním o tom, akí hlúpi sú ľudia, ktorí niečomu veria; vždy som najprv prebral danú tému, a až potom povedal: „A preto je toto hlúposť.“ Lenže v roku 2009 bola v mojej mysli otvorená otázka, či môže byť dôležité mať okolo seba ľudí, ktorí vyjadrujú pohŕdanie homeopatiou. Myslel som si, a stále si myslím, že existuje nešťastný problém, že ak sa k nejakým myšlienkam

správame zdvorilo, mnohí ľudia to na istej úrovni spracujú ako: „Ak poviem, že tomuto verím, nestane sa mi nič zlé; nestratím tým postavenie, ak poviem, že verím v homeopatiu," a že pohŕdavý smiech komikov môže pomôcť ľuďom, aby sa z tohto sna prebrali.

Dnes si myslím, že by som písal zdvorilejšie. Nezdvorilosť mala svoj účel, a myslím si, že boli ľudia, ktorým jej čítanie pomohlo; dnes však beriem vážnejšie riziko budovania spoločenstiev, kde normálnou a očakávanou reakciou na nízko postavené cudzie názory je otvorený posmech a pohŕdanie.

Napriek mojej chybe môžem s radosťou povedať, že moji čitatelia boli zatiaľ úžasne dobrí v *nepoužívaní* mojej rétoriky ako zámienky na šikanovanie alebo znevažovanie druhých. (Chcel by som tu konkrétne vyzdvihnúť Scotta Alexandra, ktorý je milším človekom než som ja, a dokáže o týchto témach písať čoraz úžasnejšie, a zrejme si zaslúži uznanie za spoluvytvorenie zdravej kultúry na *Less Wrong*.)

Ak sa môžete obzrieť a povedať, že sa vám niečo „nepodarilo", naznačuje to, že ste mali ciele. Čo som sa teda ja pokúšal urobiť?

Existuje určitý hodnotný spôsob myslenia, ktorý sa v súčasnosti ešte nevyučuje na školách. Tento konkrétny spôsob myslenia sa vôbec systematicky nevyučuje. Iba ním nasiaknu ľudia, ktorí vyrastali na knihách ako *To nemyslíte vážne, pán Feynman*, alebo ktorí mali nezvyčajne dobrého učiteľa na strednej škole.

Najslávnejší na tomto konkrétnom spôsobe myslenia je jeho vzťah s vedou a s experimentálnou metódou. Je to tá časť vedy, kde idete von a pozriete sa na svet, namiesto toho, aby ste si jednoducho niečo vymysleli. Tá časť, kde poviete „Joj" a vzdáte sa zlej teórie, keď ju experimenty nepodporia.

Ale tento konkrétny spôsob myslenia siaha ešte ďalej. Je hlbší a všeobecnejší než akési okuliare, ktoré si nasadíte pri vstupe do laboratória a pri odchode si ich zložíte. Týka sa každodenného života, hoci táto časť je jemnejšia a ťažšia. Ak však nedokážete povedať „Joj" a vzdať sa, keď to vyzerá, že niečo nefunguje, nemáte inú možnosť než si ďalej píliť pod sebou konáre. Stále musíte prikladať pílku a stále ňou musíte pohybovať. Takýchto ľudí poznáte. A niekde, v nejakom bode vášho života, na ktorý by ste najradšej nemysleli, ste aj *vy* takýmto človekom. Bolo by pekné, keby existoval nejaký spôsob myslenia, ktorý by nám mohol pomôcť s tým prestať.

Napriek tomu, aké veľké boli moje chyby, zdá sa, že tieto dva roky blogovania pomohli prekvapujúcemu množstvu ľudí prekvapujúco silne. Nefungovalo to spoľahlivo, ale občas to fungovalo.

V modernej spoločnosti sa tak málo učí o zručnostiach rozumného myslenia a rozhodovania, tak málo o matematike a fyzike za tým... že sa ukázalo, že už len prečítať si ohromnú haldu myšlienok ohľadom filozofických a vedeckých problémov môže naozaj byť prekvapujúco dobré. Prejsť tým všetkým, z tuctu rôznych uhlov, môže niekedy sprostredkovať dojem ústredného rytmu.

Pretože v konečnom dôsledku je toto celé jedna vec. Hovoril som o veľkých dôležitých vzdialených problémoch a zanedbával som okamžitý život, ale zákony, ktorými sa oboje riadi, v skutočnosti nie sú odlišné. V oblastiach, na ktoré som sa zameral, sú veľké medzery, a vyberal som si samé nesprávne príklady; ale v konečnom dôsledku je to celé jedna vec. Môžem sa hrdo obzrieť a povedať, že ešte aj po všetkých týchto chybách, ktoré som urobil, a po všetkých ďalších momentoch, keď som povedal „Joj"...

Ešte aj po piatich rokoch sa mi stále zdá, že toto je lepšie ako nič.

- Eliezer Yudkowsky, február 2015

Kniha I.

Mapa a územie

Skreslenia: Úvod

(napísal Rob Bensinger)

Nie je to tajomstvo. Z nejakého dôvodu sa to však zriedka vynára v rozhovoroch, a málo ľudí sa pýta, čo by sme s tým mali robiť. Je to skrytý vzor, nevidený za všetkými našimi úspechmi a zlyhaniami, nevidený za našimi očami. Čo je to?

Predstavte si, že siahnete do nádoby, ktorá obsahuje sedemdesiat bielych loptičiek a tridsať červených, a vytiahnete desať tajomných loptičiek. Možno tri z týchto desiatich loptičiek budú červené a vy správne uhádnete, koľko červených loptičiek je dokopy v nádobe. Ale možno sa vám podarí vybrať štyri červené loptičky, alebo nejaký iný počet. Potom sa pravdepodobne budete ohľadom celkového počtu mýliť.

Táto náhodná chyba je cenou za neúplné vedomosti, a v porovnaní s inými chybami to nie je až také zlé. Vaše odhady nebudú nesprávne *v priemere*, a čím viac sa dozviete, tým viac sa budú mať sklon zmenšovať.

Na druhej strane, predstavte si, že by biele loptičky boli ťažšie, a klesali by ku dnu danej nádoby. Potom by vaša vzorka mohla byť nereprezentatívna *konzistentným smerom*.

Tento typ chyby nazývame štatistické skreslenie. Keď je vaša metóda učenia sa o svete skreslená, dozvedieť sa viac vám možno nepomôže. Získať viac údajov môže dokonca konzistentne *zhoršiť* skreslenú predpoveď.

Ak ste naučení hlboko si vážiť vedomosti a skúmanie, toto je desivá vyhliadka. Ak si chceme byť istí, že nám ďalšie učenie pomôže, namiesto aby nám uškodilo, potrebujeme odhaliť a napraviť skreslenia v našich údajoch.

Pojem *kognitívne skreslenie* v psychológii funguje analogicky. Kognitívne skreslenie je systematická chyba v tom *ako rozmýšľame*, na rozdiel od náhodnej chyby alebo chyby spôsobenej našou nevedomosťou. Tak ako štatistické skreslenie ohýba vzorku tak, že sa menej podobá na väčšiu množinu, kognitívne skreslenia ohýbajú naše *názory* tak, že menej presne reprezentujú fakty, a ohýbajú naše *rozhodovanie* tak, že menej spoľahlivo dosahuje naše ciele.

Možno máte skreslenie optimizmu a dozviete sa, že červené loptičky možno použiť ako liek na zriedkavú tropickú chorobu, ktorá trápi vášho brata. Potom môžete preceňovať počet červených loptičiek v nádobe, pretože si *želáte*, aby loptičky boli prevažne červené. Tu nie je vaša vzorka to, čo je skreslené. To *vy* ste skreslení.

Keď však hovoríme o skreslených *ľuďoch*, musíme si dávať pozor. Zvyčajne, keď hovoríme o tom, že nejaký jednotlivec alebo skupina majú „predsudky“, robíme to, aby sme ich vyhrešili za to, že sú neféroví alebo niekomu nadŕžajú. *Kognitívne* skreslenie, to je celkom iné zvieratko. Kognitívne skreslenia sú základnou časťou toho, ako ľudia vo všeobecnosti rozmýšľajú, nie nejakou vadou, za ktorú by sme mohli viniť hroznú výchovu alebo skazenú osobnosť.[1]

Kognitívne skreslenie je systematický spôsob, ako sa vaše vrodené vzorce myslenia odchyľujú od pravdy (alebo od nejakého iného dosiahnuteľného cieľa, napríklad šťastia). Podobne ako štatistické

1 Myšlienka osobného skreslenia, mediálneho skreslenia, atď. sa podobá na štatistické skreslenie v tom, že je to *chyba*. Iné spôsoby, ako zovšeobecniť myšlienku „skreslenia„ sa namiesto toho zameriavajú jeho spojenie s nenáhodnosťou. Napríklad *induktívne* skreslenie v strojovom učení je iba množina predpokladov, ktoré učiaci sa systém použije na odvodenie predpovedí z množiny údajov. Tu je učiaci sa systém „skreslený“ v tom zmysle, že je mu ukázaný konkrétny smer; ale keďže ten smer môže byť aj *pravda*, nie je zlé, že tento konateľ má induktívne skreslenie. Je to hodnotné a nevyhnutné. Toto odlišuje induktívne „skreslenie“ pomerne jasne od iných druhov skreslenia.

skreslenia, kognitívne skreslenia môžu pokriviť náš pohľad na skutočnosť, nedajú sa vždy napraviť jednoducho tým, že zhromaždíme viac údajov, a ich účinok môže časom silnieť. Keď tým rozladeným meracím nástrojom, ktorý sa pokúšate opraviť, ste *vy* sami, odstraňovanie skreslenia je jedinečným problémom.

Stále je to však to samozrejmé miesto, kde treba začať. Ak totiž nemôžete dôverovať vlastnému mozgu, ako môžete dôverovať hocičomu inému?

Bolo by užitočné mať nejaký názov pre tento projekt prekonávania kognitívnych skreslení a prekonávania všetkých druhov chýb, kde naše mysle podkopávajú sami seba.

Mohli by sme tento projekt nazvať ľubovoľne. V tejto chvíli si však myslím, že slovo „rozumnosť“ je rovnako dobré ako hociktoré iné.

Rozumné pocity

V hollywoodskom filme, byť „rozumný“ zvyčajne znamená, že ste strnulý hyperintelektuálny stoik. Predstavte si Spocka zo seriálu *Star Trek*, ktorý „rozumne“ potláča svoje emócie, „rozumne“ sa odmieta spoliehať na intuície alebo impulzy, a ľahko ho zmätie a prekabáti hocijaký omylný alebo „nerozumný“ súper.[2]

Existuje aj celkom iné poňatie „rozumnosti“, ktoré študujú matematici, psychológovia a sociológovia. Je to viacmenej predstava *robiť to najlepšie, čo sa dá s tým, čo máte*. Rozumný človek, bez ohľadu na to, ako je ďaleko od oblasti, v ktorej sa vyzná, si vytvára najlepšie názory, aké môže mať s ohľadom na indície, ktoré má. Rozumný človek, bez ohľadu na to, v akej hroznej situácii sa ocitol, si vyberie najlepšiu možnosť na zlepšenie svoje šance na úspech.

Rozumnosť v skutočnom svete nie je ignorovanie vlastných emócií a intuície. U človeka rozumnosť často znamená začať si viac uvedomovať svoje pocity, aby ste ich mohli zohľadniť v rozhodovaní.

Rozumnosť dokonca môže byť aj vedieť, kedy nad vecami príliš *neuvažovať*. Keď si pokusné osoby vyberali, aký plagát si dajú na stenu, alebo keď predpovedali výsledok basketbalového zápasu, darilo sa im *horšie*, keď vedome analyzovali svoje dôvody.[3,4] Existujú problémy, pri ktorých nám lepšie poslúži vedomé zvažovanie, i také, pri ktorých nám lepšie poslúžia okamžité úsudky.

Psychológovia pracujúci s teóriami duálnych procesov rozlišujú v mozgu procesy „Systému 1“ (rýchle, implicitné, asociatívne, automatické poznávanie) a procesy „Systému 2“ (pomalé, explicitné, intelektuálne, ovládané poznávanie).[5] Existuje *stereotyp*, že racionalisti sa spoliehajú výlučne na Systém 2 a ignorujú svoje pocity a impulzy. Ak prekukneme tento stereotyp, niekto naozaj rozumný – kto naozaj dosahuje svoje ciele, naozaj zmierňuje škodlivé vplyvy svojich kognitívnych skreslení – sa bude výrazne spoliehať na zvyky a intuície Systému 1 tam, kde sú spoľahlivé.

Nanešťastie, samotný Systém 1 sa zdá byť *mizerným* radcom v otázke: „Kedy by som mal dôverovať Systému 1?“ Naše netrénované intuície nám nepovedia, kedy by sme sa na ne mali prestať spoliehať. Byť skreslený a neskreslený je rovnaký *pocit*.[6]

Na druhej strane, ako poznamenáva behaviorálny ekonóm Dan Ariely: sme *predvídateľne* nerozumní. Blbneme rovnakým spôsobom, znovu a znovu, systematicky.

---

2 Smutná zhoda okolností: Leonard Nimoy, herec, ktorý hral Spocka, zomrel iba pár dní pred vydaním tejto knihy. Hoci citujeme jeho postavu ako klasický príklad falošnej „hollywoodskej rozumnosti“, nechceme tým nijako znevážiť pamiatku pána Nimoya.

3 Timothy D. Wilson et al., „Introspecting About Reasons Can Reduce Post-choice Satisfaction,“ *Personality and Social Psychology Bulletin* 19 (1993): 331–331.

4 Jamin Brett Halberstadt and Gary M. Levine, „Effects of Reasons Analysis on the Accuracy of Predicting Basketball Games,“ *Journal of Applied Social Psychology* 29, no. 3 (1999): 517–530.

5 Keith E. Stanovich and Richard F. West, „Individual Differences in Reasoning: Implications for the Rationality Debate?,“ *Behavioral and Brain Sciences* 23, no. 5 (2000): 645–665, http://journals.cambridge.org/-abstract_S0140525X00003435.

6 Timothy D. Wilson, David B. Centerbar, and Nancy Brekke, „Mental Contamination and the Debiasing Problem,“ in *Heuristics and Biases: The Psychology of Intuitive Judgment,* ed. Thomas Gilovich, Dale Griffin, and Daniel Kahneman (Cambridge University Press, 2002).

Ak nedokážeme použiť svoj inštinkt na zistenie, kedy podliehame kognitívnemu skresleniu, možno stále môžeme použiť vedy o mysli.

Mnohé tváre skreslenia

Aby naše mozgy riešili problémy, vyvinuli si používanie kognitívnych heuristík – drsných skratiek, ktoré dajú správnu odpoveď často, ale nie vždy. Kognitívne skreslenia vznikajú, keď skratky použité týmito heuristikami spôsobujú pomerne konzistentné a nenápadné chyby.

Napríklad heuristika reprezentatívnosti je náš sklon odhadovať javy podľa toho, nakoľko sa zdajú reprezentatívne pre danú kategóriu. To môže viesť skresleniam ako je *klam konjunkcie*. Tversky a Kahneman→ zistili, že pokusné osoby považovali za menej pravdepodobné, že silný hráč tenisu „prehrá prvý set“ než že „prehrá prvý set, ale vyhrá zápas“.[7] Zvrátiť počiatočný neúspech sa zdá omnoho *typickejšie* pre silného hráča, takže preceňujeme pravdepodobnosť tohto komplikovaného, avšak zmysluplne znejúceho príbehu v porovnaní s pravdepodobnosťou striktne jednoduchšieho scenára.

Heuristika reprezentatívnosti môže prispievať aj k *zanedbávaniu základnej miery*, kde zakladáme svoj úsudok na tom, ako intuitívne „normálna“ je nejaká kombinácia vlastností, zanedbávajúc, aká častá je každá z týchto vlastností v základnej populácii.[8] Je pravdepodobnejšie, že Števo je hanblivý knihovník, alebo že je hanblivý predajca? Mnoho ľudí odpovedá na túto otázku rozmýšľaním, či „hanblivý“ zodpovedá ich stereotypu o týchto povolaniach. Zabúdajú vziať do úvahy, o koľko viac je predajcov než knihovníkov – v Spojených Štátoch sedemdesiatpäťkrát viac.[9]

Ďalšími príkladmi skreslenia sú *zanedbanie trvania* (vyhodnocovanie zážitkov bez ohľadu na to, ako dlho trvali), *klam utopených nákladov* (cítite sa zaviazaní k veciam, do ktorých ste investovali prostriedky v minulosti, keď by ste mali oželieť straty a ísť preč), a *sklon potvrdzovať* (berieme vážnejšie indície potvrdzujúce to, čomu už veríme).[10] [11]

Poznanie nejakého skreslenia vás však pred ním len zriedkavo ochráni. V štúdii *slepoty voči skresleniam* pokusné osoby predpovedali, že ak sa dozvedia, že nejaká maľba je dielom slávneho umelca,

---

→ http://www.econ.ucdavis.edu/faculty/nehring/teaching/econ106/readings/Extensional Versus Intuitive.pdf

7 Amos Tversky and Daniel Kahneman, „Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment,“ *Psychological Review* 90, no. 4 (1983): 293–315, doi:10.1037/0033-295X.90.4.293.

8 Richards J. Heuer, *Psychology of Intelligence Analysis* (Center for the Study of Intelligence, Central Intelligence Agency, 1999).

9 Wayne Weiten, *Psychology: Themes and Variations, Briefer Version, Eighth Edition* (Cengage Learning, 2010).

10 Raymond S. Nickerson, „Confirmation Bias: A Ubiquitous Phenomenon in Many Guises,“ *Review of General Psychology* 2, no. 2 (1998): 175.

11 *Zanedbávanie pravdepodobnosti* je ďalšie kognitívne skreslenie. V mesiacoch a rokoch nasledujúcich po útokoch z 11. septembra sa mnohí ľudia rozhodli cestovať dlhé vzdialenosti autom namiesto lietadlom. Únos nebol *pravdepodobný*, ale teraz sa zdalo, že je táto možnosť vyložená na stole; samotná možnosť únosu výrazne ovplyvnila rozhodovanie. Keď sa ľudia spoliehali na čiernobiele myslenia (autá a lietadlá sú buď „bezpečné“ alebo „nebezpečné“, bodka), v skutočnosti sa tým vystavovali omnoho väčšiemu nebezpečenstvu. Tam, kde mali zvážiť pravdepodobnosť úmrtia na ceste autom krížom cez krajinu oproti pravdepodobnosti úmrtia počas letu krížom cez krajinu – to je prvé je niekoľko stonásobne pravdepodobnejšie – sa namiesto toho spoliehali na svoj všeobecný pocit znepokojenia a úzkosti (afektívna heuristika). Rovnaký vzorec správania vidíme u detí, ktoré keď si vypočujú argumenty za a proti bezpečnostným pásom, preskakujú hore-dole medzi myšlienkami, že bezpečnostné pásy sú celkom dobrá myšlienka alebo celkom zlá, namiesto toho, aby sa pokúsili porovnať silu úvah za aj proti.

Ďalšie príklady skreslenia sú: *pravidlo vrcholu a konca* (hodnotenie udalostí v spomienkach na základe ich najsilnejšieho okamihu a toho, ako skončili); *ukotvenie* (rozhodovanie sa na základe nedávno prijatých informácií, dokonca aj keď sú irelevantné) a *sebaukotvenie* (používanie seba samého ako model pre pravdepodobné vlastnosti druhých bez dostatočného zamyslenia sa nad svojimi netypickými vlastnosťami); a *skreslenie status quo* (prílišné uprednostňovanie toho, čo je normálne a očakávané, pred tým, čo je nové a odlišné).

Cass R. Sunstein, „Probability Neglect: Emotions, Worst Cases, and Law,“ *Yale Law Journal* (2002): 61–107.

Dan Ariely, *Predictably Irrational: The Hidden Forces That Shape Our Decisions* (HarperCollins, 2008).

Boaz Keysar and Dale J. Barr, „Self-Anchoring in Conversation: Why Language Users Do Not Do What They ‚Should,‘“ in *Heuristics and Biases: The Psychology of Intuitive Judgment: The Psychology of Intuitive Judgment*, ed. Griffin Gilovich and Daniel Kahneman (New York: Cambridge University Press, 2002), 150–166, doi:10.2277/0521796792.

Scott Eidelman and Christian S. Crandall, „Bias in Favor of the Status Quo,“ *Social and Personality Psychology Compass* 6, no. 3 (2012): 270–281.

bude pre nich ťažšie neutrálne ohodnotiť kvalitu maľby. A naozaj, pokusné osoby, ktorým povedali, kto je autorom maľby, a požiadali ich vyhodnotiť jej kvalitu, vykazovali presne to skreslenie, ktoré predpovedali, v porovnaní s kontrolnou skupinou. Keď sa ich však *dodatočne* opýtali, tie isté osoby tvrdili, že ich hodnotenie maľby bolo objektívne a neovplyvnené týmto skreslením – vo všetkých skupinách![12] [13]

Obzvlášť sa vyhýbame tomu, myslieť na svoje názory ako nepresné v porovnaní s názormi druhých. Dokonca aj keď správne odhalíme skreslenia druhých, máme osobitnú *slepú škvrnu voči skresleniam*, keď dôjde na naše vlastné slabosti.[14] Pri introspekcii nevieme odhaliť žiadne „myšlienky, ktoré sa zdajú byť zaujaté“, a tak dochádzame k záveru, že musíme skrátka byť objektívnejší než všetci ostatní.[15]

Študovanie skreslení môže dokonca spôsobiť, že budete *viac* náchylní k preceňovaniu sa a sklonu potvrdzovať, keď začnete vnímať vplyv kognitívnych skreslení všade okolo vás – u každého okrem vás samotných. A slepá škvrna voči skresleniam, na rozdiel od mnohých skreslení, je *mimoriadne vážna* u ľudí, ktorí sú *mimoriadne inteligentní, premýšľaví a majú otvorenú myseľ*.[16] [17]

To je dôvod znepokojovať sa.

Ale aj tak... zdá sa, že by sme mali byť schopní robiť to lepšie. Je známe, že dokážeme zredukovať ignorovanie základnej miery tým, že myslíme na pravdepodobnosti ako na frekvencie predmetov alebo udalostí. Dokážeme minimalizovať zanedbávanie trvania tým, že trvaniu budeme venovať viac pozornosti a znázorníme si ho graficky.[18] Ľudia sa líšia v tom, ako silne vykazujú rôzne skreslenia, takže by mala existovať ešte kopa zatiaľ neznámych spôsobov ako ovplyvniť, nakoľko skreslení sme.

Ak sa však chceme zlepšiť, nestačí si listovať v zoznamoch kognitívnych chýb. Prístup k odstraňovaniu skreslení v knihe *Rozumnosť: od algoritmov po zombie* je komunikovať systematické porozumenie tomu, prečo dobré uvažovanie funguje, a ako v ňom mozog zlyháva. Do tej miery, nakoľko táto kniha plní svoj cieľ, možno jej prístup prirovnať k tomu, čo popisuje Serfas, ktorý spomína, že „roky

---

12 Katherine Hansen et al., „People Claim Objectivity After Knowingly Using Biased Strategies,“ *Personality and Social Psychology Bulletin* 40, no. 6 (2014): 691–699.

13 Podobne píše Proninová o slepote voči rodovým skresleniam:

> V jednej štúdii účastníci uvažovali o mužskom a ženskom kandidátovi na pozíciu policajného náčelníka a potom odhadovali, či je pre prácu dôležitejšia „ostrieľanosť“ alebo „formálne vzdelanie“. Výsledok bol, že účastníci uprednostňovali to pozadie, aké im bolo povedané, že má mužský kandidát (čiže ak im bolo povedané, že je „ostrieľaný“, hodnotili toto ako dôležitejšie). Účastníci boli celkom slepí voči tomuto rodovému skresleniu; dokonca čím viac si verili, že sú objektívni, tým väčšie skreslenie v skutočnosti vykazovali.

Dokonca aj keď vieme o skresleniach, poznamenáva Proninová, zostávame „naivní realisti“ ohľadom svojich vlastných názorov. Spoľahlivo sa vraciami k používaniu svojich názorov ako nedeformovaných reprezentácií toho, ako veci naozaj sú.

Eric Luis Uhlmann and Geoffrey L. Cohen, „‚I think it, therefore it’s true‘: Effects of Self-perceived Objectivity on Hiring Discrimination,“ *Organizational Behavior and Human Decision Processes* 104, no. 2 (2007): 207–223.

Emily Pronin, „How We See Ourselves and How We See Others,“ *Science* 320 (2008): 1177–1180, http://psych.princeton.edu/psychology/research/pronin/pubs/2008%20Self%20and%20Other.pdf.

14 Pri ankete 76 ľudí čakajúcich na letisku jednotlivci hodnotili seba samých ako v priemere omnoho menej podliehajúcich kognitívnym skresleniam než typický človek na letisku. Ľudia sa považujú za obzvlášť neskreslených vtedy, keď je skreslenie spoločensky nežiadúce alebo má ťažko si všimnuteľné dôsledky. Iné štúdie zistili, že ľudia, ktorí majú osobné väzby k nejakej téme vnímajú tieto väzby ako posilňujúce ich vhľad a objektivitu; ale keď vidia *iných ľudí* vykazovať *rovnaké* väzby, usudzujú, že títo ľudia sú príliš zaujatí a skreslení.

Emily Pronin, Daniel Y. Lin, and Lee Ross, „The Bias Blind Spot: Perceptions of Bias in Self versus Others,“ *Personality and Social Psychology Bulletin* 28, no. 3 (2002): 369–381.

15 Joyce Ehrlinger, Thomas Gilovich, and Lee Ross, „Peering Into the Bias Blind Spot: People’s Assessments of Bias in Themselves and Others,“ *Personality and Social Psychology Bulletin* 31, no. 5 (2005): 680–692.

16 Richard F. West, Russell J. Meserve, and Keith E. Stanovich, „Cognitive Sophistication Does Not Attenuate the Bias Blind Spot,“ *Journal of Personality and Social Psychology* 103, no. 3 (2012): 506.

17 … Nemýľme si ich s ľuďmi, ktorí si o sebe myslia, že sú mimoriadne inteligentní, premýšľaví atď. kvôli ilúzii skreslenia nadradenosti.

18 Michael J. Liersch and Craig R. M. McKenzie, „Duration Neglect by Numbers and Its Elimination by Graphs,“ *Organizational Behavior and Human Decision Processes* 108, no. 2 (2009): 303–314.

pracovnej skúsenosti v oblasti financií“ nemali vplyv na sklon ľudí podliehať klamu utopených nákladov, zatiaľ čo „počet absolvovaných kurzov účtovníctva“ pomáhal.

> V dôsledku toho môže byť potrebné rozlišovať medzi skúsenosťou a odbornosťou, kde odbornosť znamená „vyvinutie si schematického princípu, ktorý zahŕňa pojmové porozumenie problému“, čo následne umožňuje autorovi rozhodnutia rozoznať konkrétne skreslenia. Používať odbornosť ako protiopatrenie si však vyžaduje viac než iba oboznámenosť s kontextom situácie alebo byť odborníkom na konkrétnu oblasť. Vyžaduje si to, aby dotyčný plne chápal základné dôvody konkrétneho kreslenia, bol schopný všimnúť si ho v konkrétnej situácii, a aby mal poruke aj vhodné nástroje na pôsobenie proti tomuto skresleniu.[19]

Cieľom tejto knihy je položiť základy pre vytvorenie „odbornosti“ rozumnosti. To znamená nadobudnutie hlbokého porozumenia štruktúry veľmi všeobecného problému: ľudské skreslenie, sebaklam, a tisíce spôsobov, ako sofistikované myslenie dokáže poraziť samo seba.

Pár slov o tomto texte

Rozumnosť: od algoritmov po zombie začala svoj život ako séria esejí od Eliezera Yudkowskeho uverejnená v rokoch 2006 až 2009 na ekonomickom blogu *Overcoming Bias* [prekonávanie skreslenia] a z neho odvodeného komunitného blogu *Less Wrong* [menej nesprávne]. Minulý rok som pracoval s Yudkowskym v Machine Intelligence Research Institute (MIRI) [Výskumný ústav strojovej inteligencie], v mimovládke, ktorú založil v roku 2000 na štúdium teoretických požiadaviek na umelú inteligenciu (UI) múdrejšiu než človek.

Čítanie jeho textov na blogu ma priviedlo k záujmu o jeho prácu. Urobil na mňa dojem svojou schopnosťou stručne komunikovať vhľady, ktorých osvojenie mi zabralo roky štúdia analytickej filozofie. Vo svojom hľadaní, ako zmieriť anarchického a skeptického ducha vedy s rigoróznym a systematickým prístupom k bádaniu, sa Yudkowsky pokúša nielen vyvrátiť, ale aj *pochopiť* mnohé nesprávne kroky a slepé uličky, ku ktorým nás môže priviesť zlá filozofia (a zlý nedostatok filozofie). Moja nádej pri organizovaní týchto esejí do knihy je zjednodušiť ponorenie sa do nich a oceniť ich ako súvislý celok.

Výsledný úvod do rozumnosti je často osobný a neúctivý – čerpá napríklad z Yudkowskeho skúseností so svojou ortodoxnou židovskou matkou (psychiatričkou) a otcom (fyzikom), a z konverzácií na internetových diskusiách a mailových skupinách. Čitatelia, ktorý poznajú Yudkowskeho fikciu *Harry Potter a metódy rozumnosti*→, jeho vedecky orientovanú verziu J. K. Rowlingovej kníh *Harry Potter*, nájdu to isté neúctivé obrazoborectvo a mnohé z rovnakých ústredných pojmov.

Štylisticky tvoria eseje v tejto knihe spektrum od „živej učebnice“ cez „zbierku zaujímavých myšlienok“ až po „burcujúci manifest“, a podobne pestrý je aj obsah. *Rozumnosť: od algoritmov po zombie* zhŕňa stovky Yudkowskeho článkov z blogu do dvadsiatich šiestich „postupností“, kapitol tvorených tematicky prepojenými článkami. Samotné tieto postupnosti sú zoskupené do šiestich kníh, zahŕňajúcich nasledujúce témy:

Kniha 1 - **Mapa a územie**. Čo je to názor, a prečo niektoré názory fungujú lepšie než iné? Tieto štyri postupnosti vysvetľujú *bayesovské* pojmy rozumnosti, názoru a indície. Opakujúca sa téma: veci, ktoré nazývame „vysvetlenia“ alebo „teórie“ nemusia vždy fungovať ako *mapy* na navigovanie vo svete. Dôsledkom je riziko, že si pomiešame svoje myšlienkové mapy s inými predmetmi v našej sade nástrojov.

Kniha 2 - **Ako naozaj zmeniť svoj názor**. Táto vec pravda vyzerá pomerne užitočne. Prečo teda stále skáčeme k záverom, zakopávame sa do zákopov, a opakujeme tie isté chyby? Prečo sme takí *slabí* v nadobúdaní presných názorov, a ako to môžeme robiť lepšie? Týchto sedem postupností preberá motivované rozmýšľanie a sklon potvrdzovať, s osobitným zameraním na ťažko viditeľné druhy sebaklamu a na pascu „používania argumentov ako vojakov“.

19 Sebastian Serfas, *Cognitive Biases in the Capital Investment Context: Theoretical Considerations and Empirical Experiments on Violations of Normative Rationality* (Springer, 2010).

→ http://hpmor.com/

Kniha 3 - **Stroj v duchovi**. Prečo sme sa *nevyvinuli* ako rozumnejší? Ešte aj keď zohľadníme obmedzené zdroje, zdá sa, že by sme mohli dostávať omnoho viac epistemickej muziky za svoje indíciové peniaze. Aby sme dostali realistický obraz toho, ako a prečo naše mysle vykonávajú svoje biologické funkcie, potrebujeme nazrieť pod kapotu a uvidieť presnejšie, ako funguje evolúcia, a ako fungujú naše mozgy. Tieto tri sekvencie znázorňujú, ako sa dokonca aj filozofi a vedci môžu mýliť, keď sa spoliehajú na intuitívne, netechnické evolučné alebo psychologické zdôvodnenia. Ak umiestnime svoje mysle do širšieho priestoru systémov orientovaných na ciele, môžeme odhaliť niektoré osobitosti ľudského myslenia a nahliadnuť, ako takéto systémy dokážu „stratiť svoj cieľ".

Kniha 4 - **Obyčajná skutočnosť**. V akom druhu sveta žijeme? Aké je naše miesto v tomto svete? Stavajúc na príkladoch predchádzajúcich sekvencií o tom, ako fungujú evolučné a kognitívne modely, týchto šesť postupností skúma prirodzenosť mysle a povahu fyzikálneho zákona. Popri aplikovaní a zovšeobecňovaní minulých lekcií o vedeckých tajomstvách a stručnosti, tieto eseje nastoľujú nové otázky o tom, akú rolu by veda mala hrať v rozumnosti jednotlivca.

Kniha 5 - **Obyčajné dobro**. Čo robí niečo *hodnotným* - morálne, esteticky, alebo rozvážne? Tieto tri postupnosti sa pýtajú, ako môžeme zdôvodniť, upraviť a zakotviť v prírode naše hodnoty a túžby. Cieľom je nájsť spôsob, ako pochopiť naše ciele bez kompromitovania nášho úsilia naozaj ich dosiahnuť. Tu je najväčším problémom vedieť, kedy dôverovať svojim zamotaným, zložitým, na konkrétnych prípadoch závisiacim impulzom o tom, čo je správne a čo nesprávne, a kedy ich nahradiť jednoduchšími princípmi bez výnimiek.

Kniha 6 - **Stať sa silnejším**. Ako môžu jednotlivci a spoločenstvá uplatniť toto všetko v praxi? Tieto tri postupnosti začínajú autobiografickým výkladom Yudkowskeho vlastných najväčších filozofických prehmatov, s jeho radami, ako si myslí, že by to iní mohli robiť lepšie. Kniha končí odporúčaniami vyvinúť učebné osnovy rozumnosti založené na dôkazoch, a vytvoriť skupiny a organizácie na podporu študentov, učiteľov, výskumníkov, a priateľov, ktorí o to majú záujem.

Tieto postupnosti dopĺňajú „medzihry", eseje z Yudkowskeho osobnej webovej stránky http://www.yudkowsky.net. Tieto rôznymi spôsobmi nadväzujú na postupnosti, napríklad Dvanásť cností rozumnosti[→] poeticky zhŕňa mnohé lekcie z knihy *Rozumnosť: od algoritmov po zombie*, a ďalšie eseje z nej často citujú.

Kliknutie na hviezdičku na spodku eseje vás privedie na jej pôvodnú [anglickú] verziu na stránke *Less Wrong* (kde môžete zanechať komentáre) alebo na Yudkowskeho webovú stránku. Na internete môžete nájsť aj [anglický] slovník pojmov z knihy *Rozumnosť: od algoritmov po zombie* na adrese http://wiki.lesswrong.com/wiki/RAZ_Glossary.

Mapa a územie

Táto prvá kniha začína postupnosťou o kognitívnom skreslení: „Predvídateľné chyby[→]". Zvyšok knihy nebude iba o tejto téme; zlé zvyky a zlé myšlienky sú dôležité, aj keď sú výsledkom obsahu našich myslí, nielen jej štruktúry. Takto vzniknuté a vymyslené chyby budú znázornené v nasledujúcich postupnostiach, začínajúc diskusiou v postupnosti „Falošné názory[→]" o spôsoboch, ako sa očakávania človeka môžu odchyľovať od jeho vyhlasovaných názorov.

Rozprávanie o nerozumnosti by nebolo úplné, keby neposkytlo nejakú teóriu o tom, ako funguje *rozumnosť* - alebo keby sa jeho „teória" skladala iba z hmlistých banalít, bez presného vysvetľujúceho mechanizmu. Postupnosť „Všímať si zmätok[→]" sa pýta, prečo je užitočné založiť svoje správanie na „rozumných" očakávaniach, a aký je to pocit, keď to robíte.

„Tajomné odpovede[→]" sa následne pýta, či veda rieši tieto problémy za nás. Vedci zakladajú svoje modely na opakovateľných experimentoch, nie na špekuláciách ani klebetách. A veda má skvelé skóre

---

→ Kapitola Medzihra: Dvanásť cností rozumnosti, strana 659

→ Časť A: Predvídateľné chyby, strana 20

→ Časť B: Falošné názory, strana 36

→ Časť C: Všímať si zmätok, strana 50

→ Časť D: Tajomné odpovede, strana 64

v porovnaní s anekdotami, náboženstvom, a... prakticky všetkým ostatným. Potrebujeme sa stále báť „falošných„ názorov, sklonu potvrdzovať, skreslenia spätného pohľadu, a podobných vecí, keď pracujeme so spoločenstvom ľudí, ktorí chcú javy *vysvetľovať*, nielen rozprávať pôsobivé príbehy?

Po tomto nasleduje Jednoduchá pravda→, samostatná alegória o podstate vedomosti a názoru.

Kognitívne skreslenie je však to, čo poskytuje najjasnejší a najpriamejší pohľad do vecí v našej psychiky, do tvaru našich heuristík a logiky našich obmedzení. Preto začneme skreslením.

Existuje citát v proto-taoistickom texte *Čuang'C*, ktorý hovorí: „Pasca na ryby existuje kvôli rybe; keď už si rybu chytil, môžeš na pascu zabudnúť.“[20]

Pozývam vás, aby ste túto knihu skúmali v tomto duchu. Používajte ju, ako by ste používali pascu na ryby, stále pamätajúc na účel, ktorý pre ňu máte. Odneste si to, čo dokážete použiť, dokiaľ je to pre vás užitočné; zvyšok zahoďte. Kiež vám vaše ciele dobre poslúžia.

Poďakovania

Som ohromne vďačný nasledujúcim osobám: Nate Soares, Elizabeth Tarleton, Paul Crowley, Brienne Strohl, Adam Freese, Helen Toner, a tuctom dobrovoľníkov za korektúry častí tejto knihy. Osobitne úprimne ďakujem Alexovi Vermeerovi, ktorý doviedol túto knihu do jej dokončenia, a Tsvi Benson-Tilsenovi, ktorý prečesal celú túto knihu, aby zaručil jej čitateľnosť a súdržnosť.

→ Kapitola Medzihra: Jednoduchá pravda, strana 86

20 Zhuangzi and Burton Watson, *The Complete Works of Zhuangzi* (Columbia University Press, 1968).

A: Predvídateľné chyby

1. Čo myslím slovom „rozumnosť“?

Myslím tým nasledujúce:

1. **Epistemická racionalita** (rozumné poznanie): systematické zvyšovanie správnosti vašich názorov.
2. **Inštrumentálna racionalita** (rozumné konanie): systematické dosahovanie svojich hodnôt.

Keď otvoríte oči a pozriete sa na miestnosť okolo vás, zistíte polohu svojho počítača voči stolu, a zistíte polohu svojej police na knihy voči stene. Ak sa niečo pokazilo vo vašich očiach alebo vo vašom mozgu, potom váš myšlienkový model môže hovoriť, že existuje polica tam, kde žiadna polica neexistuje, a keď si prídete po knihu, budete sklamaní.

Také je mať nepravdivý názor, mapu sveta, ktorá nezodpovedá územiu. Epistemická rozumnosť je, že si namiesto toho budujeme presné mapy. Tento súlad medzi názorom a skutočnosťou sa ľudovo nazýva „pravda→“ a ja tento názov ochotne používam.

Na druhej strane, inštrumentálna rozumnosť je, že skutočnosť *kormidlujete* – posielate budúcnosť tam, kam chcete, aby išla. Je to umenie vyberania si činov, ktoré vedú v výsledkom hodnoteným vyššie na vašom rebríčku preferencií. Toto občas nazývam „víťazstvo“.

Rozumnosť sa teda týka vytvárania pravdivých názorov a robenia víťazných rozhodnutí.

Hľadanie „pravdy“ tu neznamená odmietanie neistých alebo nepriamych indícií. Pozeranie na izbu okolo vás a budovanie si jej myšlienkovej mapy sa v princípe neodlišuje od názoru, že Zem má tekuté jadro, alebo že Július Cézar bol holohlavý. Tieto otázky, ktoré sú od vás vzdialené v priestore a čase, môžu vyzerať ľahkomyseľnejšie a abstraktnejšie než otázky o vaše polici na knihy. Napriek tomu existujú fakty v záležitosti stavu zemského jadra v roku 2015 n.l. a stavu Cézarovej hlavy v roku 50 p.n.l. Niekto fakty môžu mať na vás reálne účinky aj vtedy, keď nikdy nestretnete Cézara ani zemské jadro zoči-voči.

A „víťazstvo“ tu nemusí prichádzať na úkor druhých. Životný projekt sa môže týkať spolupráce alebo sebaobetovania namiesto súťaže. „Vaše hodnoty“ tu znamenajú *čokoľvek, na čom vám záleží*, vrátane druhých ľudí. Nie je to obmedzené na *sebecké* alebo *nie spoločné* hodnoty.

Keď ľudia povedia: „X je rozumné!“, zvyčajne je to iba dôraznejší spôsob ako povedať: „Myslím si, že X je pravda“ alebo: „Myslím si, že X je dobré.“ Načo je nám teda ďalšie slovo „rozumný“, keď už máme slová „pravdivý“ a „dobrý“?

Analogický argument by sme mohli dať proti používaniu slova „pravda“. Netreba hovoriť „je pravda, že sneh je biely“, keď môžete jednoducho povedať „sneh je biely“. Čo robí myšlienku pravdy užitočnou je, že nám dovoľuje hovoriť o všeobecných vlastnostiach súladu medzi mapou a územím. „Pravdivé modely zvyčajne dávajú lepšie experimentálne predpovede než nepravdivé modely“ je užitočné zovšeobecnenie, a nedá sa urobiť bez použitia pojmov ako „pravda“ alebo „presný“.

Podobne: „Rozumní konatelia robia rozhodnutia, ktoré maximalizujú pravdepodobnostné očakávanie koherentnej funkcie úžitku“ je ten druh myšlienky, ktorý závisí na pojme (inštrumentálnej) rozumnosti, zatiaľ čo: „Je rozumné jesť zeleninu“ pravdepodobne možno nahradiť: „Je užitočné jesť zeleninu“ alebo: „Je vo vašom záujme jesť zeleninu.“ Pojem ako je „rozumný“ potrebujeme, aby sme vedeli poznamenať všeobecné fakty o tých spôsoboch myslenia, ktoré systematicky vytvárajú pravdu alebo hodnotu – a o tých systematických spôsoboch, ako za týmito štandardmi zaostávame.

Niekedy experimentálni psychológovia odhalia ľudské uvažovanie, ktoré vyzerá veľmi čudne. Napríklad→ niekto hodnotí pravdepodobnosť výroku „Bill hrá džez“ ako *nižšiu* než pravdepodobnosť výroku „Bill je účtovník, ktorý hrá džez“. To vyzerá ako čudný úsudok, keďže ľubovoľný konkrétny účtovník hrajúci džez je zrejme hráčom džezu. Na akú vyššiu moc sa však odvolávame tvrdením, že tento úsudok je *nesprávny*?

→ Kapitola Medzihra: Jednoduchá pravda, strana 86

→ http://lesswrong.com/lw/ji/conjunction_fallacy/

Experimentálni psychológovia používajú dva zlaté štandardy: *teóriu pravdepodobnosti* a *teóriu rozhodovania*.

Teória pravdepodobnosti je množina zákonov, ktorými sa riadia rozumné názory. Matematika pravdepodobnosti rovnakým spôsobom popisuje (a) ako zistiť, kde je vaša polica na knihy, (b) ako zistiť teplotu zemského jadra, a (c) ako odhadnúť, koľko vlasov bolo na hlave Júlia Cézara. Toto všetko je ten istý problém, ako pracovať indície a pozorovania, aby sme upravili („aktualizovali“) svoje názory. Podobne, teória rozhodovania je množina zákonov, ktorými sa riadi rozumné konanie, a dá sa rovnako použiť bez ohľadu na to, aké má človek ciele a dostupné možnosti.

Nech „P(niečo)“ znamená „pravdepodobnosť, že sa niečo stane“, a „P(A, B)“ znamená „pravdepodobnosť, že sa stane aj A, aj B“. Keďže je všeobecným zákonom teórie pravdepodobnosti, že P(A) ≥ P(A, B), úsudok, že P(„Bill hráva džez“) je menej ako P(„Bill hráva džez“ & „Bill je účtovník“) označujeme ako nesprávny.

Z technického hľadiska by ste povedali, že takýto pravdepodobnostný úsudok je *nebayesovský*. Názory, ktoré sú v súlade s koherentnou distribúciou pravdepodobnosti a rozhodnutia, ktoré maximalizujú očakávanú pravdepodobnosť koherentnej funkcie úžitku, sa nazývajú „bayesovské“.

Všimnite si, že moderné poňatie rozumnosti sa netýka slovného uvažovania. Ako príklad som uviedol otvorenie očí, obzrenie sa, a vytvorenie si myšlienkového modelu izby obsahujúcej policu s knihami pri stene. Moderné poňatie rozumnosti je dostatočne všeobecné na to, aby zahrnulo vaše oči a zrakové oblasti vášho mozgu ako veci-ktoré-mapujú. Rovnako dobre zahŕňa aj vašu intuíciu bez slov. Matematika sa nestará o to, čo používame to isté slovo, „rozumný“, pri rozprávaní o Spockovi aj pri rozprávaní o bayesovstve. Matematika modeluje dobré spôsoby dosahovania cieľov alebo mapovania sveta, bez ohľadu na to, či sú tieto spôsoby v súlade s našimi predsudkami a stereotypmi o tom, aká by mala byť „rozumnosť“.

Problém, čo v praxi označovať ako „rozumné“, sme týmto celkom nevyčerpali, najmä z dvoch dôvodov:

Po prvé, bayesovské formalizmy vo svojej plnej podobe sú pre väčšinu problémov zo skutočného života výpočtovo nezvládnuteľné. Nikto nedokáže *naozaj* spočítať všetku tú matematiku a riadiť sa ňou, rovnako ako nedokážete predpovedať pohyb cien na burze vypočítaním pohybu kvarkov.

Preto existuje celá internetová stránka s názvom „LessWrong“ [menej chybne] namiesto jednej strany, ktorá by jednoducho uviedla formálne axiómy a basta. Existuje ešte celé umenie hľadania pravdy a dosahovania hodnôt *zvnútra ľudskej mysle*: musíme spoznať svoje vlastné chyby, prekonať svoje skreslenia, vyvarovať sa sebaklamu, dostať sa do dobrého emocionálneho stavu, aby sme čelili pravde a urobili, čo treba, atď. atď. a tak ďalej.

Po druhé, niekedy sa spochybňuje význam samotnej matematiky. Presné pravidlá teórie pravdepodobnosti sú spochybňované napríklad antropickými problémami→, v ktorých je nejasný počet pozorovateľov. Presné pravidlá teórie rozhodovania sú spochybňované napríklad problémami ako je Newcombov→291, v ktorých iní aktéri môžu predpovedať vaše rozhodnutie skôr než ho urobíte.[21]

V takýchto prípadoch je márne pokúšať sa problém uzavrieť vymyslením nejakej novej definície→154 slova „rozumný“ a hovoriť: „A preto moja obľúbená odpoveď, *podľa definície* →172, je to, čomu sa hovorí ‚rozumné‘.“ To jednoducho vyvolá otázku, prečo by niekto mal venovať pozornosť vašej definícii. Ja sa nezaujímam o teóriu pravdepodobnosti kvôli tomu, že je to sväté písmo, ktoré nám priniesol Laplace. Zaujímam sa o aktualizovanie názorov bayesovským spôsobom (s occamovskými

---

→ http://www.anthropic-principle.com/primer.html

→291 Kapitola 291. Newcombov problém a ľutovanie rozumnosti, strana 654

21 Poznámka redaktora: Dobrý úvod k Newcombovmu problému je Holt. Všeobecnejšie, definície a vysvetlenia mnohých z pojmov v tejto knihe nájdete na webovej stránke http://wiki.lesswrong.com/wiki/RAZ_Glossary.
Jim Holt, „Thinking Inside the Boxes,“ *Slate* (2002), http://www.slate.com/articles/arts/egghead/2002/02/-thinkinginside_the_boxes.html.

→154 Kapitola 154. Podobenstvo o bolehlave, strana 290

→172 Kapitola 172. Argumentovanie „podľa definície“, strana 323

pôvodnými pravdepodobnosťami[→25]), pretože očakávam, že takéto myslenie nás systematicky približuje, veď viete, k *správnosti*, k mape, ktorá odráža územie.

A potom existujú otázky „Ako myslieť“, ktoré sa nezdajú dostatočne zodpovedané ani teóriou pravdepodobnosti, ani teóriou rozhodovania – napríklad otázka, aký mať pocit z pravdy, keď už ju máte[→2]. Aj tu, pokúsiť sa definovať „rozumnosť“ konkrétnym spôsobom nezdôvodňuje odpoveď, iba ju podsúva.

Nejdem sa tu hádať o význam slova[→163], ani keď je tým slovom „rozumnosť“. Zmyslom priraďovania postupností písmen ku konkrétnym pojmom je umožniť dvom ľuďom *komunikovať*[→165] – pomôcť dopraviť myšlienky z jednej mysle do druhej. Pomocou manipulácie, ktoré významy patria ku ktorým slovám, nemôžete zmeniť skutočnosť ani dokázať myšlienku.

Ak teda rozumiete, na ktorý pojem sa *zhruba odvolávam* týmto slovom „rozumnosť“, a na ktoré podvýrazmi „rozumné myslenie“ a „rozumné konanie“, potom *sme komunikovali*: dosiahli sme všetko, čo sa dalo dosiahnuť rozprávaním, ako definovať „rozumnosť“. Čo nám ešte zostáva na diskusiu, nie je *aký význam* priradíme jednotlivým slabikám „ro–zum–nosť“; zostáva nám diskutovať o tom, *ako dobre myslieť*.

Ak poviete: „Rozumný (epistemicky) názor je X, ale pravdivý názor je Y,“ potom pravdepodobne používate slovo „rozumný“ na označenie niečoho iného než na čo myslím ja. (Napríklad „rozumnosť“ by mala byť *konzistentná po úvahe* - „rozumne“ sa pozrieť na indíciu a „rozumne“ zvážiť, ako vaša myseľ spracováva túto indíciu, by nemalo viesť k dvom rôznym záverom.)

Podobne, ak zistíte, že hovoríte: „Rozumné (inštrumentálne) je urobiť X, ale správna vec je urobiť Y,“ potom takmer určite používate niektoré zo slov „rozumné“ alebo „správne“ spôsobom, s ktorým nesúhlasím. Ja používam slovo „rozumný“ *normatívne*, aby som poukázal na žiadúce spôsoby myslenia.

V takomto prípade – alebo v hociktorom inom prípade, kde hrozí spor – by ste mali dosadiť konkrétnejšiu reč[→167] namiesto „rozumného“: „Sebe by som najviac prospel útekom preč, ale dúfam, že by som sa aspoň pokúsil odtiahnuť to dievča z koľajníc“ alebo „Kauzálna teória rozhodovania je zvyčajne formulovaná tak, že by ste podľa nej mali vziať dve krabice v Newcombovom probléme[→291], ale ja by som radšej mal milión dolárov.“

V skutočnosti odporúčam, aby ste si túto esej prečítali ešte raz, ale každé slovo „rozumný“ nahraďte slovom „onaký“ a sledujte, či to nejako mení konotácie toho, čo hovorím. Ak áno, potom hovorím: nesnažte sa rozumnosť, ale o onakosť.

Slovo „rozumný“ má svoje možné úskalia, ale je kopec *nie* hraničných prípadov, kde slovo „rozumný“ veľmi dobre *komunikuje*, o čo človeku ide; podobne aj „nerozumný“. V takýchto prípadoch sa ho nebojím používať.

Človek by si však mal dať pozor, aby to slovo nepoužíval *príliš*. Nedostávame žiadne body za to, že ho iba nahlas vyslovujeme. Ak priveľa hovoríte o Ceste, nedosiahnete ju.

*→

2. Rozumné cítenie

O „rozumnosti“ ľudia radi hovoria, že odporuje všetkým emóciám - že všetky naše žiale a všetky naše radosti sú automaticky proti logike, pretože sú to *city*. Pritom je dosť čudné, že v teórii pravdepodobnosti neviem nájsť žiadnu vetu dokazujúcu, že by som sa mal tváriť ľadovo chladne a bez výrazu.

→25 Kapitola 25. Occamova britva, strana 56
→2 Kapitola 2. Rozumné cítenie, strana 22
→163 Kapitola 163. Debaty o definíciách, strana 307
→165 Kapitola 165. Argumentovanie bežným používaním, strana 312
→167 Kapitola 167. Hrajte so slovami tabu, strana 315
→291 Kapitola 291. Newcombov problém a ľutovanie rozumnosti, strana 654
→ http://lesswrong.com/lw/31/what_do_we_mean_by_rationality/

Je teda rozumnosť nezávislá od cítenia? Nie; naše emócie vyplývajú z našich modelov skutočnosti. Ak budem veriť, že môjho mŕtveho brata našli živého, budem šťastný; ak sa zobudím a uvedomím si, že to bol sen, budem smutný. P. C. Hodgell povedal: „Čo môže byť zničené pravdou, nech sa zničí.“ Šťastie môjho snívajúceho ja odporovalo pravde. Môj smútok po zobudení je rozumný; nie je žiadna pravda, ktorá by ho zničila.

Rozumnosť začína tým, že sa pýtame, aký je svet, ale šíri sa ako vírus do každej ďalšej myšlienky, ktorá závisí na tom, aký si myslíme, že svet je. Keď hovorím o vašich názoroch na to, „aký je svet“, myslím tým hocičo, čo veríte, že existuje v skutočnosti, hocičo, čo buď existuje alebo neexistuje, ľubovoľný prvok množiny „vecí, ktoré môžu spôsobiť iné veci.“ Ak veríte, že vo vašej skrini je škriatok, ktorý vám zväzuje šnúrky na topánkach, potom je to názor na to, aký je svet. Vaše topánky sú skutočné – môžete ich zdvihnúť. Ak existuje niečo, čo môže chytiť a zviazať vaše šnúrky, aj to musí byť skutočné, časť rozsiahlej siete príčin a následkov, ktorú voláme „vesmír“.

Hnevať sa na škriatka, ktorý vám zviazal šnúrky, zahŕňa stav mysle, ktorý nie je *iba* o tom, aký je svet. Predpokladajme, že ako buddhistu, alebo pacienta po lobotómii, alebo skrátka veľmi flegmatického človeka by vás zistenie, že máte zviazané šnúrky, nenahnevalo. Neovplyvňovalo by to vaše očakávania o svete – stále by ste očakávali, že otvoríte skriňu a nájdete šnúrky zviazané. Váš hnev ani pokoj by tu nemali ovplyvniť váš odhad, lebo čo sa deje vo vašej skrini, nezávisí od emocionálneho stavu mysle; aj keď môže byť náročné myslieť takto jasne.

Ale pocit hnevu je spojený so stavom mysle, ktorý *je* o tom, aký je svet; hneváte sa, *pretože* si myslíte, že vám škriatok zviazal šnúrky. Kritérium rozumnosti sa šíri ako vírus, od počiatočnej otázky, či vám škriatok zviazal alebo nezviazal šnúrky, až do výsledného hnevu.

Stať sa rozumnejším – získavať lepšie odhady o tom, aký je svet – môže pocity zmierniť *alebo zosilniť*. Niekedy utekáme pred silnými citmi tým, že popierame fakty, že cúvame pred pohľadom na svet, ktorý viedol k tejto silnej emócii. Ak je to tak, potom ako budete študovať zručnosti rozumnosti a cvičiť sa v nepopieraní faktov, vaše city zosilnejú.

V začiatkoch som si nikdy nebol celkom istý, či je *v poriadku* cítiť niečo silne – či je to dovolené, či je to správne. Nemyslím si, že tento zmätok vznikol len z môjho mladistvého nechápania rozumnosti. Všimol som si podobné ťažkosti u ľudí, ktorí sa ani nepokúšajú byť racionalistami; keď sú šťastní, pochybujú, či naozaj smú byť šťastní, a keď sú smutní, nie sú si celkom istí, či pred touto emóciou utekať alebo nie. Prinajmenšom od Sokratových čias, a asi aj dávno pred ním, ste mohli pôsobiť vzdelane a sofistikovane tak, že ste nikdy pred nikým nedali najavo, že vám na niečo veľmi záleží. Cítiť je *hanba* – v slušnej spoločnosti sa to skrátka nerobí. Mali by ste vidieť, ako čudne sa na mňa ľudia pozerajú, keď si uvedomia, ako veľmi mi záleží na rozumnosti. Nemyslím si, že je to kvôli nezvyčajnej téme, ale že nie sú zvyknutí vidieť duševne zdravých dospelých ľudí, ktorým na *niečom* viditeľne záleží.

Teraz však viem, že na silných pocitoch nie je nič zlé. Odkedy som si osvojil pravidlo: „Čo môže byť zničené pravdou, nech sa zničí“, uvedomil som si aj: „Čo pravda živí, nech prekvitá.“ Keď sa stane niečo dobré, som šťastný, a nemám v hlave zmätok, či je rozumné, aby som bol šťastný. Keď sa stane niečo hrozné→, neutekám pred smútkom hľadaním falošnej útechy a falošnej nádeje. Predstavujem si minulosť a budúcnosť ľudstva, desiatky miliárd smrtí počas dejín, biedu a strach, hľadanie odpovedí, trasúce sa ruky siahajúce nahor z toľkej krvi, čím by sme sa jedného dňa mohli stať, až z hviezd urobíme svoje mestá, všetku tú temnotu a všetko to svetlo – viem, že to nikdy nedokážem naozaj pochopiť, a že to neviem povedať slovami. Napriek celej svojej filozofii sa stále hanbím priznávať silné pocity, a vám je pravdepodobne nepríjemné počúvať o nich. Ale teraz už viem, že cítiť je rozumné.

* →

→ http://yudkowsky.net/yehuda.html

→ http://lesswrong.com/lw/hp/feeling_rational/

3. Načo pravda? A...

Niektoré komentáre na *Overcoming Bias* sa týkali otázky, prečo by sme mali hľadať pravdu. (Našťastie väčšina nespochybňovala, čo to je pravda→.) Naša motivácia nastaviť svoje myšlienky na rozumnosť, ktorá určuje, či je dané nastavenie „dobré“ alebo „zlé“, sa odvíja z dôvodu, prečo vlastne tú pravdu chceme hľadať.

Je napísané: „Prvou cnosťou je zvedavosť.“ Zvedavosť je jedným dôvodom hľadať pravdu, a hoci nie je jediným dôvodom, má istú zvláštnu a obdivuhodnú čistotu. Ak je vaším motívom zvedavosť, budete otázky uprednostňovať podľa toho, ako tieto otázky samotné šteklia váš osobný estetický zmysel. Zamotanejšia otázka s väčšou pravdepodobnosťou neúspechu môže byť hodna viac úsilia než jednoduchšia, skrátka preto, lebo je zábavnejšia.

Ako som napísal, ľudia si často predstavujú rozum a emócie ako súperov. Keďže zvedavosť je emócia, tuším, že niektorí ľudia budú namietať voči považovaniu zvedavosti za časť rozumnosti. Ja osobne označujem emóciu za „nie rozumnú“ vtedy, keď sa opiera o mylné názory, alebo presnejšie, o poznávacie správanie vytvárajúce omyly: „Ak sa k tvojej tvári blíži železo, o ktorom veríš, že je horúce, avšak ono je chladné, Cesta odporuje tvojmu strachu. Ak sa k tvojej tvári blíži železo, o ktorom veríš, že je chladné, avšak ono je horúce, Cesta odporuje tvojmu pokoju.“ A naopak, emócia vyvolaná správnymi názormi alebo rozumnými poznávacími úvahami je „rozumná emócia“; čo má tú výhodu, že aj pokoj môžeme považovať za emocionálny stav a nie za privilegovaný štandard.

Keď si ľudia myslia, že „emócie“ a „rozum“ sú protiklady, mám podozrenie, že v skutočnosti myslia na Systém 1 a Systém 2 – rýchle úsudky založené na vnímaní verzus pomalé úsudky založené na uvažovaní. Úsudky založené na uvažovaní nie sú vždy pravdivé, ani úsudky založené na vnímaní nie sú vždy nepravdivé; je teda veľmi dôležité rozoznávať toto rozdelenie od „rozumnosti“. Oba systémy môžu slúžiť pravde, alebo proti nej bojovať, podľa toho, ako sa použijú.

Okrem čírej emocionálnej zvedavosti, aké ďalšie môžu byť motívy túžiť po pravde? No, možno chcete dosiahnuť nejaký konkrétny cieľ v skutočnom svete, napríklad postaviť lietadlo, a preto potrebujete vedieť nejakú konkrétnu pravdu o aerodynamike. Alebo prostejšie, chcete čokoládové mlieko a preto chcete vedieť, či v miestnych potravinách majú čokoládové mlieko, aby ste sa rozhodli, či pôjdete tam alebo niekam inam. Ak chcete pravdu kvôli tomuto, potom budete otázky uprednostňovať podľa očakávaného úžitku ich informácie – nakoľko prípadné odpovede ovplyvnia vaše rozhodnutia, nakoľko na vašich rozhodnutiach záleží, nakoľko očakávate, že nájdete odpoveď, ktorá zmení vaše rozhodnutia oproti terajším.

Hľadať pravdu iba pre jej praktické využitie sa môže zdať neslušné – nemali by sme túžiť po pravde iba kvôli nej samotnej? – avšak takéto hľadanie je mimoriadne dôležité, pretože vytvára vonkajšie overovacie kritérium: ak sa vaše lietadlo zrúti z oblohy, alebo ak pôjdete do obchodu a nebudú mať čokoládové mlieko, je to znamenie, že ste niečo spravili zle. Dostávate spätnú väzbu, ktoré spôsoby myslenia fungujú, a ktoré nie. Čistá zvedavosť je nádherná vec, ale nemusí vydržať dosť dlho na to, aby overila svoje odpovede, keď už príťažlivosť tajomstva pominula. Zvedavosť ako ľudská emócia tu bola už dávno pred starovekými Grékmi. Na cestu Vedy však ľudstvo pevne postrčilo uvedomenie, že niektoré spôsoby myslenia odhaľujú názory, pomocou ktorých môžeme *meniť svet*. Čo sa týka čistej zvedavosti, túto túžbu rovnako dobre uspokojovalo aj rozprávanie o bohoch a hrdinoch pri táboráku, a nikto si neuvedomil, že je na tom niečo zle.

Existujú aj motívy hľadania pravdy iné ako zvedavosť a pragmatizmus? Tretím dôvodom, ktorý mi napadá, je morálka: Veríte, že hľadať pravdu je vznešené a dôležité a hodnotné. Hoci aj takýto ideál pripisuje pravde vnútornú hodnotu, tento stav mysle sa veľmi odlišuje od zvedavosti. Byť zvedavý, čo je za oponou, nie je rovnaký pocit ako veriť, že máte morálnu povinnosť tam pozrieť. V tomto druhom stave mysle máte väčší sklon veriť, že aj niekto *iný* by sa mal pozrieť za oponu, prípadne ho karhať, ak úmyselne zavrie oči. Preto označujem ako „morálku“ názor, že hľadanie pravdy je pragmaticky dôležité „pre spoločnosť“, a preto je to povinnosť nás všetkých. Pri tejto motivácii budete otázky

→ Kapitola Medzihra: Jednoduchá pravda, strana 86

uprednostňovať podľa vašich ideálov o tom, ktoré pravdy sú najdôležitejšie (nie najužitočnejšie ani najzaujímavejšie); alebo podľa vašich morálnych ideálov o tom, kedy a za akých okolností je povinnosť hľadať pravdu najsilnejšia.

Mám sklon podozrievať morálku ako motiváciu k rozumnosti, *nie* preto, že by som odmietal tento morálny ideál, ale preto, že privoláva určité problémy. Je príliš ľahké osvojiť si, ako naučené morálne povinnosti, spôsoby myslenia, ktoré sú hroznými prešľapmi pri tanci. Predstavte si pána Spocka z seriálu *Star Trek*, naivný archetyp rozumnosti. Spockov emocionálny stav je vždy nastavený na „pokoj“, aj keď je to celkom neprimerané. Často udáva veľa platných číslic pri pravdepodobnostiach, ktoré sú hrubo nekalibrované. (Napríklad: „Kapitán, ak zamierite s Enterprise priamo do tejto čiernej diery, naša pravdepodobnosť prežitia je iba 2,234%.“ Napriek tomu, v deviatich prípadov z desiatich Enterprise vyviazne. Aký tragický blázon udáva štyri platné číslice pre údaj, ktorý je o dva rády mimo?) Napriek tomu si podľa tohto známeho obrazu mnoho ľudí predstavuje povinnosť byť „rozumný“ – nečudo, že sa s ňou vrelo nestotožňujú. Urobiť z rozumnosti morálnu povinnosť znamená dať jej všetky hrozné stupne voľnosti svojvoľného kmeňového zvyku. Ľudia dôjdu k nesprávnej odpovedi, a potom rozhorčene namietajú, že konali správne, namiesto aby sa zo svojej chyby poučili.

A predsa, ak máme *zlepšiť* svoje zručnosti rozumnosti, prekročiť výkonové štandardy lovcov a zberačov, potrebujeme rozvážne názory na to, ako správne myslieť. Keď si pre seba píšeme nové myšlienkové programy, začínajú v Systéme 2, systéme rozvážnosti, a až pomaly – ak vôbec – sa cvičením dostanú do nervových obvodov riadiacich Systém 1. Ak teda existujú určité druhy myslenia, o ktorých zistíme, že sa im chceme *vyhnúť* – napríklad skreslenia – musia byť na úrovni Systému 2 reprezentované ako zákazy myslieť takýmto spôsobom; ako deklarovaná povinnosť vyhýbania sa.

Ak chceme poznať pravdu, najúčinnejšie to dosiahneme myslením určitými spôsobmi namiesto iných; a toto sú techniky rozumnosti. Niektoré z týchto techník rozumnosti zahŕňajú prekonávanie určitého druhu prekážok, skreslenia...

*→

4. ...čo je teda skreslenie?

Skreslenie je určitý druh prekážky na ceste k nášmu cieľu dosiahnuť pravdu – považujeme ho za „prekážku“, pretože máme pravdu ako cieľ – ale existujú aj mnohé prekážky, ktoré nie sú „skreslenia“.

Keby sme rovno začali otázkou: „Čo je to skreslenie?“, išli by sme na to z nesprávnej strany. Ako hovorí príslovie: „Existuje štyridsať druhov bláznovstva, ale iba jeden druh zdravého rozumu.“ Pravda je úzky terč, malá oblasť konfiguračného priestoru, ktorú chceme zasiahnuť. „Ľúbi ma alebo neľúbi?“ môže byť binárna otázka, ale „ $E=mc^2$ “ je malá bodka v priestore všetkých rovníc, tak ako vyhrávajúci žreb lotérie v priestore všetkých žrebov. Chyba nie je výnimočná situácia; práve úspech je a priori taký nepravdepodobný, že si vyžaduje vysvetlenie.

Nezačíname morálnou povinnosťou „znížiť skreslenia“, pretože skreslenia sú škaredé a zlé a Skrátka To Nerobíme. S takýmto zmýšľaním môže človek skončiť, ak nadobudol deontologickú povinnosť „rozumnosti“ spoločenskou osmózou, ktorá vedie ľudí k snahe vykonávať techniky bez docenenia ich dôvodov. (A práve toto je škaredé a zlé a Skrátka To Nerobíme, podľa knihy *To nemyslíte vážne, pán Feynman*, ktorú som čítal ako dieťa.)

My skôr chceme dosiahnuť pravdu, z hocijakého dôvodu, a na ceste k svojmu cieľu nachádzame rôzne prekážky. Tieto prekážky si nie sú navzájom celkom nepodobné – existujú napríklad prekážky súvisiace s nedostatkom dostupnej výpočtovej sily alebo s vysokou cenou informácií. Zhodou okolností jedna veľká skupina prekážok vyzerá, že má určitú spoločnú povahu – tvorí zhluk v priestore prekážok na ceste k pravde – a tento zhluk sme nazvali „skreslenia“.

Čo je to skreslenie? Vieme sa pozrieť na tento empirický zhluk a nájsť presný test členstva? Možno zistíme, že to naozaj nevieme vysvetliť lepšie, než ukázať na pár reprezentatívnych príkladov a dúfať,

→ http://lesswrong.com/lw/go/why_truth_and/

že poslucháč porozumie. Ak ste vedec, ktorý práve začína skúmať oheň, môže byť omnoho múdrejšie ukázať na táborák a povedať: „Oheň je tamtá oranžovosvetlá horúca vec," namiesto povedania: „Definujem oheň ako alchymickú transmutáciu substancií, ktoré uvoľňujú flogiston." Nemali by ste niečo ignorovať iba preto, že to neviete definovať. Ja nedokážem spamäti zacitovať rovnice všeobecnej teórie relativity, ale napriek tomu, ak skočím z útesu, spadnem. A to isté môžeme povedať o skresleniach – nedopadnú na nás o nič menej tvrdo, ak sa ukáže, že nevieme presne definovať, čo to „skreslenie" vlastne je. Môžeme teda ukázať na omyl konjunkcie, prehnanú dôveru, heuristiku reprezentatívnosti, ignorovanie základnej miery a povedať: „Takéto veci."

S prihliadnutím na horeuvedené, zdá sa, že označujeme ako „skreslenia" tie prekážky na ceste k pravde, ktoré nespôsobuje cena informácie, ani obmedzená výpočtová sila, ale tvar nášho vlastného myšlienkového mechanizmu. Tento mechanizmus je napríklad evolučne optimalizovaný na ciele, ktoré aktívne bránia presnosti poznania; napríklad mechanizmus na vyhrávanie hádok prispôsobený na politické situácie. Alebo výberový tlak presnosť poznania odkláňa; napríklad veríme tomu, čomu veria druhí, aby sme spoločensky zapadli. Alebo, pri klasických heuristikách a skresleniach, mechanizmus funguje podľa známeho algoritmu, ktorý prináša nejaký úžitok, ale produkuje aj systematické chyby: samotná heuristika dostupnosti nie je skreslenie, ale spôsobuje známe, presne opísateľné skreslenia. Naše mozgy robia niečo nesprávne, a po mnohých pokusoch a náročnom premýšľaní niekto opíše tento problém spôsobom, ktorému Systém 2 dokáže porozumieť; potom to nazývame „skreslenie". Aj keby sme nemali žiaden lepší spôsob poznávania, stále je to chyba, ktorú pomenovateľným spôsobom vytvára konkrétny druh poznávacieho mechanizmu – nie pretože toho mechanizmu máme primálo, ale kvôli samotnému tvaru tohto mechanizmu.

„Skreslenia" odlišujeme od chýb vychádzajúcich z poznávacieho obsahu, ako sú prijaté názory alebo prijaté morálne povinnosti. Tie nazývame skôr „chyby" než „skreslenia", a je omnoho ľahšie ich napraviť, keď si ich na sebe všimneme. (Zdrojom tejto chyby, alebo zdrojom zdroja tejto chyby, však v konečnom dôsledku môže byť nejaké skreslenie.)

„Skreslenia" odlišujeme od chýb spôsobených poškodením jednotlivého ľudského mozgu, alebo prijatými kultúrnymi normami; skreslenia spôsobuje mechanizmus, ktorý majú všetci ľudia spoločný.

Platón nemal „skreslenie", keď nepoznal teóriu relativity – nemal túto informáciu ako získať, jeho nevedomosť nespôsoboval tvar jeho myšlienkového mechanizmu. Ale ak Platón veril, že z filozofov by boli lepší králi, pretože on sám bol filozof – a ak tento názor vznikol vďaka všeobecnému politickému inštinktu prispôsobenému na sebapresadzovanie, a nie preto, lebo Platónovi jeho otecko hovoril, že každý má morálnu povinnosť presadzovať vládu svojej profesie, ani preto, lebo Platón v detstve čuchal priveľa lepidla – potom to bolo skreslenie, či už na to Platóna niekedy niekto upozornil alebo nie.

Nemusí byť ľahké napraviť skreslenia. Možno sa ani napraviť nedajú. Ale keď sa pozrieme na svoj vlastný myšlienkový mechanizmus a vidíme kauzálne vysvetlenie pomenovateľnej triedy chýb; a keď to vyzerá, že tento problém spôsobuje vyvinutý tvar tohto mechanizmu, a nedostatok tohto mechanizmu alebo jeho konkrétny zlý obsah; potom to nazývame skreslenie.

Ja osobne vidím našu snahu ako nadobúdanie osobných zručností rozumnosti, ako zlepšovanie techniky hľadania pravdy. Úlohou je dosiahnuť pozitívny cieľ pravdy, nie vyhnúť sa negatívnemu cieľu chyby. Priestor chýb je rozsiahly, nekonečné množstvá nekonečne rozmanitých chýb. Je ťažké opísať taký obrovský priestor: „Čo je pravdou o jednom jablku, nemusí byť pravdou o druhom jablku; preto sa o jednom jablku dá povedať viac než o všetkých jablkách na svete." Priestor úspechu je užší, preto sa o ňom dá viac povedať.

Aj keď (ako vidíte) nie som proti diskusii o definíciách, mali by sme si pamätať, že to nie je náš prvoradý cieľ Sme tu preto, aby sme pokračovali vo veľkej ľudskej snahe o pravdu: pretože vedomosti zúfalo potrebujeme, a navyše, sme zvedaví. Pre tento cieľ sa usilujme prekonávať rôzne prekážky na ceste, či už ich voláme „skreslenia" alebo nie.

*→

5. Dostupnosť

Heuristika dostupnosti je posudzovanie frekvencie alebo pravdepodobnosti udalosti podľa toho, ako ľahko si vieme spomenúť na jej príklady.

Známa štúdia z roku 1978 od Lichtensteina, Slovica, Fischhoffa, Laymana a Combsa, „Odhadovaná frekvencia smrteľných udalostí“, študovala chyby pri vyčísľovaní vážnosti rizík alebo odhadovaní, ktoré z dvoch nebezpečenstiev nastáva častejšie.[22] Pokusné osoby si mysleli, že nehody spôsobujú zhruba rovnako veľa úmrtí ako choroby; mysleli si, že vražda je častejšou príčinou smrti než samovražda. V skutočnosti choroby spôsobujú asi 16-krát viac smrtí ako nehody a samovražda je dvakrát častejšia než vražda.

Zrejmá hypotéza pre tieto skreslené názory je, že o vraždách sa hovorí viac ako o samovraždách – preto si človek skôr spomenie, že počul o vražde, než o samovražde. Nehody sú dramatickejšie než choroby – možno preto si ich ľudia skôr zapamätajú, alebo si na ne skôr spomenú. V roku 1979 nasledujúca štúdia od Combsa a Slovica ukázala, že skreslené odhady pravdepodobnosti silno korelujú (0,85 a 0,89) so skreslenými frekvenciami správ v dvoch novinách.[23] To nerieši, či si na vraždy ľahšie spomíname, pretože sa o nich viac píše, alebo či o vraždách noviny viac píšu, pretože sú farbistejšie (a teda sa lepšie zapamätajú). Ale každopádne tu účinkuje skreslenie dostupnosti. Selektívne spravodajstvo je jedným z hlavných zdrojov skreslení dostupnosti. V pravekom prostredí ste väčšinu z toho, o čom ste vedeli, zažili osobne; alebo ste o tom počuli priamo od súkmeňovca, ktorý to videl. Medzi vami a danou udalosťou bola zvyčajne nanajvýš jedna vrstva selektívneho spravodajstva. Na dnešnom internete môžete vidieť správy, ktoré cestou k vám prešli rukami šiestich blogerov – šesť postupných filtrov. V porovnaní s našimi predkami žijeme vo väčšom svete, v ktorom sa omnoho viac stane a omnoho menšia časť z toho dôjde k nám – omnoho silnejší efekt výberu, ktorý môže vytvárať omnoho väčšie skreslenia dostupnosti.

V skutočnom živote je nepravdepodobné, že niekedy stretnete Billa Gatesa. Vďaka selektívnemu spravodajstvu médií však môžete byť v pokušení porovnávať svoje životné úspechy s ním – a utrpieť príslušnú hedonistickú pokutu. Objektívna frekvencia Billa Gatesa je 0,000 000 000 15, ale počujete o ňom omnoho častejšie. Na druhej strane, 19 % ľudí na tejto planéte žije z menej než jedného dolára denne, ale pochybujem, že každý piaty článok, ktorý čítate na internete, napísali oni.

Používanie dostupnosti asi vytvára skreslenie absurdity→; nespomínate si na udalosti, ktoré sa nestali, a preto ich pravdepodobnosť považujete za nulovú. Ak v poslednej dobe nenastali záplavy (ale pravdepodobnosti sa stále dajú pomerne ľahko vypočítať), ľudia si odmietajú kupovať poistenie proti záplavám dokonca aj keď je silno dotované a jeho cena je hlboko pod skutočnou poistnou hodnotou. Kunreuther a kolektív (1993) naznačuje, že za podceňovanie hrozby záplavy môže „neschopnosť jednotlivcov predstaviť si záplavy, ktoré sa nikdy nestali... Ľudia na záplavových územiach sú do veľkej miery väzňami svojej skúsenosti... Nedávne záplavy zrejme nastavujú hornú hranicu výšky škody, o ktorej si manažeri myslia, že by ich mala znepokojovať.“[24]

Burton a kol. (1978) uvádza, že keď sa postavia priehrady a protizáplavové bariéry, zníži sa frekvencia záplav a tým sa zrejme vytvorí falošný pocit bezpečnosti, ktorý vedie k zníženej opatrnosti.[25]

→ http://lesswrong.com/lw/gp/whats_a_bias_again/

22 Sarah Lichtenstein et al., „Judged Frequency of Lethal Events,“ [Vnímaná frekvencia smrteľných udalostí] *Journal of Experimental Psychology: Human Learning and Memory* 4, no. 6 (1978): 551–578, doi:10.1037/0278-7393.4.6.551.

23 Barbara Combs and Paul Slovic, „Newspaper Coverage of Causes of Death,“ [Novinové spravodajstvo a príčiny smrti] *Journalism & Mass Communication Quarterly* 56, no. 4 (1979): 837–849, doi:10.1177/107769907905600420.

→ http://lesswrong.com/lw/j4/absurdity_heuristic_absurdity_bias/

24 Howard Kunreuther, Robin Hogarth, and Jacqueline Meszaros, „Insurer Ambiguity and Market Failure,“ [Neistota poisťovateľa a zlyhanie trhu] *Journal of Risk and Uncertainty* 7 (1 1993): 71–87, doi:10.1007/BF01065315.

25 Ian Burton, Robert W. Kates, and Gilbert F. White, *The Environment as Hazard*, [Životné prostredie ako nebezpečenstvo] 1st ed. (New York: Oxford University Press, 1978).

Hoci postavenie priehrad znižuje *frekvenciu* záplav, škoda *pripadajúca na jednu záplavu* je potom o toľko väčšia, že priemerné ročné škody *narastú*. Múdry človek by zo spomienky na malé riziká extrapoloval možnosť veľkých rizík. Namiesto toho sa zdá, že zážitok malých rizík nastavuje vnímanú hornú hranicu rizika. Spoločnosť dobre chránená proti malým nebezpečenstvám nekoná žiadne akcie proti veľkým nebezpečenstvám, keď sú pravidelné malé záplavy odstránené, začne sa na záplavových rovinách stavať. Spoločnosť vystavená pravidelným malým nebezpečenstvám berie tieto malé nebezpečenstvá ako hornú hranicu nebezpečenstva, chráni sa proti pravidelným malým záplavám, ale nie proti občasným veľkým.

Pamäť nie je vždy dobrým radcom ohľadom pravdepodobností v minulosti, a už vôbec nie v budúcnosti.

*→

6. Priťažujúce podrobnosti

> Iba potvrdzujúci detail, s cieľom dať umeleckú dôveryhodnosť inak nudnému a nepresvedčivému rozprávaniu…
>
> --Pooh-Bah v opere *Mikado* od Gilberta a Sullivana[26]

Klam konjunkcie→ je, keď ľudia hodnotia pravdepodobnosť P(A, B) ako vyššiu než pravdepodobnosť P(B), napriek vete, že P(A, B) ≤ P(B). Napríklad v jednom experimente z roku 1981, 68 % pokusných osôb hodnotilo ako pravdepodobnejšie, že „Reagan poskytne federálnu podporu slobodným matkám a zoškrtá federálnu podporu lokálnym vládam“ než že „Reagan poskytne federálnu podporu slobodným matkám“.

Dlhá séria múdro navrhnutých pokusov, ktoré vyvrátili alternatívne hypotézy a definitívne potvrdili štandardnú interpretáciu→, potvrdila, že klam konjunkcie nastáva preto, lebo „dosadzujeme hodnotenie reprezentatívnosti za hodnotenie pravdepodobnosti“. Pridaním ďalších podrobností môžete dosiahnuť, že výsledok bude vyzerať *typickejší* pre proces, ktorý ho vytvára. Tvrdenie, že Reagan podporí slobodné matky, môžete urobiť dôveryhodnejším, keď *pridáte* tvrdenie, že Reagan *ešte aj* zoškrtá podporu lokálnym vládam. Nedôveryhodnosť jedného tvrdenia sa vyváži dôveryhodnosťou druhého tvrdenia; „spriemerujú sa“.

Čiže: Pridanie podrobnosti môže urobiť scenár NAPOHĽAD DÔVERYHODNEJŠÍ, hoci sa tým celková udalosť nevyhnutne STANE MENEJ PRAVDEPODOBNOU.

Ak je to tak, potom by *čisto hypoteticky* mali existovať futuristi, ktorí splietajú nehorázne dôveryhodné a podrobné dejiny budúcnosti, alebo ľudia, ktorí zhltnú obrovské balíky nepodložených tvrdení, ak sa zostavia okolo niekoľkých presvedčivo znejúcich tvrdení. Ak vám predložia klam konjunkcie v nahom, priamom porovnaní, možno sa vám pri tomto konkrétnom probléme podarí vedome sa opraviť. Ale to je iba nalepenie náplaste na problém, nie jeho oprava vo všeobecnosti.

V pokuse z roku 1982, kde profesionálni predpovedači systematicky priraďovali vyššie pravdepodobnosti „Rusko napadne Poľsko, s následným prerušením diplomatických vzťahov medzi USA a ZSSR“ oproti „prerušenie diplomatických vzťahov medzi USA a ZSSR“, každá pokusná skupina dostala iba jeden z výrokov.[27] Akú stratégiu mohli použiť títo predpovedači, ako skupina, aby odstránili klam konjunkcie, keď žiaden z nich jednotlivo priamo o tomto porovnaní nevedel? Keď žiaden z nich ani nevedel, že tento pokus bude *o klame* konjunkcie? Ako boli zlepšiť svoje odhady pravdepodobnosti?

Náplasť na jeden špeciálny chyták nerieši problém vo všeobecnosti. Chyták je príznakom, nie chorobou. Bolo by to rovnako hlúpe ako, povedzme, zakazovať vreckové nožíky v lietadlách.

→ http://lesswrong.com/lw/j5/availability/

26 William S. Gilbert and Arthur Sullivan, *The Mikado*, Opera, 1885.

→ http://lesswrong.com/lw/ji/conjunction_fallacy/

→ http://lesswrong.com/lw/jj/conjunction_controversy_or_how_they_nail_it_down/

27 Tversky and Kahneman, „Extensional Versus Intuitive Reasoning.“

Čo mohli predpovedači urobiť, aby sa vyhli klamu konjunkcie, aj keď porovnanie priamo nevideli, ani nevedeli, že ich niekto ide skúšať z klamu konjunkcie? Zdá sa mi, že by si museli všimnúť slovo „a“. Museli by si naň dávať pozor – nielen dávať pozor, ale uskočiť pred ním. Aj keby nevedeli, že ich potom výskumníci budú skúšať konkrétne z klamu konjunkcie. Museli by si všimnúť konjunkciu *dvoch detailov* a byť *šokovaní* tou drzosťou, že ich niekto žiada o podporu takejto šialene komplikovanej predpovede. A museli by takú pravdepodobnosť *podstatne* penalizovať – aspoň o štyri rády, podľa detailov experimentu.

Možno by pomohlo, keby sa predpovedači zamysleli nad možnými dôvodmi, prečo by USA a Sovietsky Zväz mohli prerušiť diplomatické vzťahy. Scenár nie je „USA a Sovietsky Zväz náhle bezdôvodne prerušia diplomatické vzťahy“, ale „USA a Sovietsky Zväz prerušia diplomatické vzťahy z nejakého dôvodu“.

A čo pokusné osoby, ktoré hodnotili „Reagan poskytne federálnu podporu slobodným matkám a zoškrtá federálnu podporu lokálnym vládam“? Opäť, mali byť šokované slovom „a“. Navyše, mali absurdity *sčítavať* – absurdita nech je logaritmus pravdepodobnosti, takže sa dá sčítať – a nie priemerovať. Mali si pomyslieť: „Reagan môže ale nemusí zoškrtať podporu lokálnym vládam (1 bit), ale vyzerá veľmi nepravdepodobne, že by podporil slobodné matky (4 bity). *Celková* absurdita: 5 bitov.“ Alebo trebárs: „Reagan by nepodporil slobodné matky. Celé to neprichádza do úvahy. Ten druhý výrok to už len zhoršuje.“

Podobne, predstavte si šesťstennú kocku so štyrmi zelenými a dvoma červenými stenami. Pokusné osoby mali staviť na jednu z postupností (1) „ČZČČČ“, (2) „ZČZČČČ“ alebo (3) „ZČČČČČ“, či sa objaví niekde medzi 20 hodmi kocky.[28] 65% pokusných osôb si vybralo „ZČZČČČ“, ktoré je jednoznačne horšie ako „ČZČČČ“, pretože ľubovoľná postupnosť obsahujúca „ZČZČČČ“ obsahuje aj „ČZČČČ“. Ako mohli pokusné osoby konať lepšie? Všimnúť si, že jedna postupnosť obsahuje druhú? Možno; ale to je iba náplasť, ktorá nerieši základný problém. Presným výpočtom pravdepodobností? To by základný problém určite vyriešilo, ale nie vždy môžete pravdepodobnosť presne vypočítať.

Pokusné osoby heuristicky prehrali tým, že si mysleli: „Aha! Postupnosť 2 má vyšší pomer zelených k červeným! Mal by som staviť na postupnosť 2!“ Aby heuristicky vyhrali, pokusné osoby by si mali myslieť: „Aha! Postupnosť 1 je *krátka*! Mal by som staviť na postupnosť 1!“

Mali by cítiť silnejší *emocionálny dopad* Occamovej britvy – cítiť *každú* pridanú podrobnosť ako bremeno, dokonca aj jeden hod kockou navyše.

Kedysi som sa rozprával s niekým, kto bol očarený nejakým neopatrným futuristom. (Takým, ktorý pridával množstvo dobre znejúcich detailov.) Pokúšal som sa vysvetliť, prečo nie som rovnako zhypnotizovaný týmito úžasnými, neuveriteľnými teóriami. Vysvetlil som teda klam konjunkcie, konkrétne ten pokus „prerušenie vzťahov ± napadnutie Poľska“. On na to: „Dobre, ale ako to súvisí s...“ Ja: „Je pravdepodobnejšie, že sa vesmíry replikujú *z nejakého dôvodu*, než že sa replikujú *pomocou čiernych dier, pretože vyspelé civilizácie stavajú čierne diery, pretože vesmíry sa vyvíjajú tak, aby to robili.*“ On: „Ach.“

Až dovtedy necítil tieto podrobnosti navyše ako bremeno navyše. Naopak, to boli potvrdzujúce detaily, dodávajúce dôveryhodnosť rozprávaniu. Niekto vám predloží balík zvláštnych myšlienok z ktorých *jedna* je, že vesmíry sa replikujú. Potom predloží podporu *pre tvrdenie, že vesmíry sa replikujú*. To však nie je podpora pre ten balík[→], aj keď to celé podáva ako jeden príbeh.

Musíte tieto podrobnosti rozmotať. Musíte chytiť každú z nich samostatne a opýtať sa: „Ako vieme *túto* podrobnosť?“ Niekto vám vykreslí obrázok úpadku ľudstva do nanotechnologickej vojny, kde Čína odmietne dodržiavať medzinárodné dohody o kontrole, nasledujú preteky vo vývoji zbraní... Počkaj chvíľu – ako vieš, že to bude Čína? Máš tam vo vrecku krištáľovú guľu, alebo sa len tešíš, že si futurista? Odkiaľ pochádzajú všetky tieto podrobnosti? Odkiaľ pochádza *táto konkrétna* podrobnosť?

28 Amos Tversky and Daniel Kahneman, „Judgments of and by Representativeness,“ in *Judgment Under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky (New York: Cambridge University Press, 1982), 84–98.

→ http://en.wikipedia.org/wiki/Package-deal_fallacy

Pretože je napísané:

> *Ak môžeš uľahčiť svoje bremeno, musíš to urobiť.*
>
> *Niet takej slamky, ktorá by nemala silu zlomiť ti chrbát.*

⁂→

7. Klam plánovania

Medzinárodné letisko v Denveri→ otvorili o 16 mesiacov neskôr a jeho cena prekročila pôvodný plán o 2 miliardy dolárov (videl som aj údaj 3,1 miliardy dolárov). Eurofighter Typhoon→, spoločný vojenský projekt niekoľkých európskych krajín, dokončili o 54 mesiacov neskôr za cenu 19 miliárd dolárov namiesto 7 miliárd. Dom opery v Sydney→ je asi najznámejšia oneskorená stavba všetkých čias, pôvodne odhadovaný na rok 1963 za 7 miliónov dolárov, naozaj dokončený v roku 1973 za 102 miliónov dolárov.[29]

Sú to izolované pohromy, ktorým venujeme pozornosť vďaka selektívnej dostupnosti→5? Sú to príznaky nesprávne motivovanej byrokracie alebo vlády? Áno, veľmi pravdepodobne. Ale existuje aj príslušné kognitívne skreslenie opakovane potvrdené pokusmi, v ktorých plánovali jednotlivci.

Buehler a kol. (1995) požiadal svojich študentov o odhad, kedy si (študenti) myslia, že dokončia svoje osobné akademické projekty.[30] Konkrétne sa výskumníci pýtali na odhad času, kedy si študenti myslia, že ich osobný projekt bude hotový s pravdepodobnosťou 50%, 75% a 99%. Chcete hádať, koľko študentov stihlo skončiť do vlastného odhadu úrovne pravdepodobnosti 50%, 75% a 99%?

- 13% pokusných osôb dokončilo svoj projekt v čase, ktorému priradili pravdepodobnosť 50%;
- 19% dokončilo v čase, ktorému priradili pravdepodobnosť 75%;
- a iba 45% (menej ako polovica!) dokončilo v čase odhadovanej pravdepodobnosti 99%.

Ako píše Buehler a kol. (2002), „Výsledky pre pravdepodobnosť 99% sú osobitne šokujúce: Napriek požiadavke, aby urobili vysoko konzervatívny odhad, predpoveď, ktorú prakticky naisto cítia, že ju splnia, dôvera študentov vo svoje časové odhady vysoko prevyšovala ich výsledky.“[31]

Všeobecnejšie sa tento jav nazýva „klam plánovania“. Klam plánovania je, keď si ľudia myslia, že vedia plánovať, ha ha.

Návod k problému algoritmu plánovania objavil Newby-Clark a kol. (2000), ktorý zistil, že:

- Keď požiadame pokusné osoby o predpovede založené na realistickom „najpravdepodobnejšom“ scenári; alebo
- Keď požiadame pokusné osoby o vysnívaný „najlepší možný“ scenár...

…získané výsledky od seba *nemožno rozoznať*.[32]

→ http://lesswrong.com/lw/jk/burdensome_details/

→ http://en.wikipedia.org/wiki/Denver_International_Airport

→ http://en.wikipedia.org/wiki/Eurofighter_Typhoon

→ http://en.wikipedia.org/wiki/Sydney_Opera_House

29 Roger Buehler, Dale Griffin, and Michael Ross, „Inside the Planning Fallacy: The Causes and Consequences of Optimistic Time Predictions,“ in Gilovich, Griffin, and Kahneman, *Heuristics and Biases*, 250–270.

→5 Kapitola 5. Dostupnosť, strana 27

30 Roger Buehler, Dale Griffin, and Michael Ross, „Exploring the ‚Planning Fallacy‘: Why People Underestimate Their Task Completion Times,“ *Journal of Personality and Social Psychology* 67, no. 3 (1994): 366–381, doi:10.1037/0022-3514.67.3.366; Roger Buehler, Dale Griffin, and Michael Ross, „It’s About Time: Optimistic Predictions in Work and Love,“ *European Review of Social Psychology* 6, no. 1 (1995): 1–32, doi:10.1080/14792779343000112.

31 Buehler, Griffin, and Ross, „Inside the Planning Fallacy.“

32 Ian R. Newby-Clark et al., „People Focus on Optimistic Scenarios and Disregard Pessimistic Scenarios While Predicting Task Completion Times,“ *Journal of Experimental Psychology: Applied* 6, no. 3 (2000): 171–182, doi:10.1037/1076-898X.6.3.171.

Keď ľudí žiadame o „realistický“ scenár, predstavia si, že všetko pôjde presne podľa plánu, žiadne *nečakané* meškania ani *nepredvídané* katastrofy – rovnaká predstava ako ich „najlepší možný prípad“.

Ako sa ukazuje, skutočnosť zvyčajne dáva výsledky o čosi horšie než „najhorší možný prípad“.

Na rozdiel od väčšiny kognitívnych skreslení, pri klame plánovania poznáme dobrú heuristiku na odstránenie skreslenia. Nebude fungovať na chaos rozmerov medzinárodného letiska v Denveri, ale bude fungovať na mnohé osobné plány a dokonca aj na nejaké organizačné veci malého rozsahu. Jednoducho použije „pohľad zvonka“ namiesto „pohľadu zvnútra“.

Ľudia majú sklon zostavovať svoje predpovede tak, že pomyslia na čiastkové jedinečné vlastnosti danej úlohy, a vytvoria si scenár, ako túto úlohu zamýšľajú dokončiť – čo je to, čo si zvyčajne predstavujeme ako *plánovanie*. Keď chcete niečo urobiť, musíte si naplánovať kde, kedy, ako; zistiť, koľko času a koľko prostriedkov potrebujete; predstaviť si kroky od začiatku po úspešný záver. Toto všetko je „pohľad zvnútra“, neberie do úvahy nečakané zdržania a nepredvídané katastrofy. Ako sme už videli, požiadať ľudí, aby si predstavili „najhorší možný prípad“ stále nestačí na prekonanie ich optimizmu – nepredstavujú si dosť Murphyho zákona.

Vonkajší pohľad je, keď sa úmyselne *vyhýbate* mysleniu na osobitné jedinečné vlastnosti daného projektu, a iba sa opýtate, ako dlho trvalo dokončiť *zhruba* podobné projekty v minulosti. Je to proti intuícii, pretože pohľad zvnútra má omnoho viac detailov – je tu pokušenie myslieť si, že starostlivo na mieru ušitá predpoveď, ktorá zohľadní všetky dostupné údaje, dá lepšie výsledky.

Pokusy však ukazujú, že čím podrobnejšia je predstava pokusných osôb, tým optimistickejšie (a menej presné) sa stávajú. Buehler a kol. (2002) požiadal skupinu pokusných osôb, aby popísali veľmi konkrétne plány svojich vianočných nákupov – kde, kedy a ako.[33] Táto skupina v priemere očakávala, že dokončí nakupovanie viac ako týždeň pred Vianocami. Druhej skupiny sa jednoducho opýtal, kedy očakávajú, že dokončia svoje vianočné nákupy, a priemerná odpoveď bola 4 dni. Obe skupiny skončili v priemere 3 dni pred Vianocami.

Podobne, Buehler a kol. (2002) v opise medzikultúrnej štúdie zistili, že japonskí študenti očakávali, že svoje eseje dokončia 10 dní pred termínom. V skutočnosti ich dokončili 1 deň pred termínom. Keď sa ich pýtali, kedy dokončovali podobné úlohy v minulosti, odpovedali: „jeden deň pred termínom“.[34] Toto je sila pohľadu zvonka oproti pohľadu zvnútra.

Podobne sa zistilo, že skúsení ľudia mimo organizácie, ktorí poznajú menej podrobností, ale môžu sa oprieť o relevantné spomienky, sú často omnoho menej optimistickí a omnoho presnejší než skutoční plánovači a realizátori.

Existuje teda dosť spoľahlivý spôsob, ako opraviť klam plánovania, ak robíte niečo, čo sa *zhruba* podobá na referenčnú triedu predchádzajúcich projektov. Skrátka sa opýtajte, ako dlho trvali podobné projekty v minulosti, a nezohľadňujte pritom *žiadne* osobitné vlastnosti tohto projektu. Ešte lepšie, opýtajte sa skúseného človeka mimo organizácie, ako dlho trvali podobné projekty.

Dostanete odpoveď, ktorá bude znieť príšerne dlho, a bude jasné, že neodráža pochopenie špeciálnych dôvodov, prečo táto konkrétna úloha zaberie menej času. Tá odpoveď je pravdivá. Vyrovnajte sa s tým.

*→

8. Ilúzia priehľadnosti: Prečo vám nikto nerozumie

Pri skreslení spätného pohľadu→ ľudia, ktorí poznajú výsledok nejakej situácie, veria, že tento výsledok sa dal ľahko predpovedať. Keď už výsledok poznáme, prerozprávame si celú situáciu→[33]

33 Buehler, Griffin, and Ross, „Inside the Planning Fallacy.“

34 Tamtiež.

→ http://lesswrong.com/lw/jg/planning_fallacy/

→ http://lesswrong.com/lw/il/hindsight_bias/

→33 Kapitola 33. Falošná kauzalita, strana 67

vo svetle tohto výsledku. Aj keď nás na to upozornia, nedokážeme si toto vysvetlenie odmyslieť natoľko, aby sme vžili do kože niekoho, kto nevie to, čo my.

S tým úzko súvisí *ilúzia priehľadnosti*. Vždy vieme, čo myslíme *svojimi* vlastnými slovami, a preto očakávame, že to budú vedieť aj druhí. Keď čítame, čo sme sami napísali, ľahko sa nám text správne vysvetľuje, keď vieme, čo sme tým naozaj mysleli. Ťažké je vžiť sa do kože niekoho, kto musí vysvetľovať naslepo, riadený iba slovami.

June odporučí Markovi reštauráciu; Mark sa tam navečeria a zistí, že (a) jedlo nie je nič moc a obsluha je podpriemerná (b) jedlo je vynikajúce a obsluha bezchybná. Potom Mark zanechá June na odkazovači nasledujúcu správu: „June, práve som dovečeral v reštaurácii, ktorú si mi odporučila, a musím povedať, že bola úžasná, fakt úžasná.“ Keysar (1994) predložil skupine pokusných osôb scenár (a) a 59 % považovalo Markovu správu za sarkastickú *a mysleli si, že June by tento sarkazmus vnímala*.[35] Iné pokusné osoby dostali scenár (b) a iba 3 % z nich si mysleli, že June bude vnímať Markovi správu ako sarkastickú. Keysar a Barr (2002) naznačujú, že pokusné osoby si vypočuli skutočnú správu zo záznamu.[36] Keysar (1998) ukázal, že ak pokusným osobám povieme, že reštaurácia bola hrozná, *ale Mark chcel zaprieť svoju reakciu*, pokusné osoby veria, že June nebude vnímať sarkazmus v (tej istej) správe:[37]

> To, že si [June] všimne sarkazmus, predpovedali s rovnakou pravdepodobnosťou, keď sa [Mark] pokúšal skryť svoju negatívnu skúsenosť, ako keď mal pozitívnu skúsenosť a bol úprimný. Účastníci teda brali Markov *komunikačný zámer* ako priehľadný. Akoby predpokladali, že June bude vnímať ten zámer, ktorý Mark chcel, aby vnímala.[38]

„Hus visí vysoko“ je archaické anglické slovné spojenie, ktorý sa v modernom jazyku nepoužíva. Keysar a Bly (1995) povedali jednej skupine pokusných osôb, že „hus visí vysoko“ znamená, že budúcnosť vyzerá dobre; druhej skupine pokusných osôb povedali, že „hus visí vysoko“ znamená, že budúcnosť vyzerá pochmúrne.[39] Pokusných osôb sa potom pýtali, ktorý z týchto dvoch významov by *neinformovaný* poslucháč skôr pripísal tomuto slovnému spojeniu. Každá skupina si myslela, že poslucháč by chápal toto slovné spojenie „štandardne“.

(Ďalšie testované slovné spojenia boli „išiel na niekoho ako ujo“, „ísť pozdĺž paluby“ a „vyložiť do levandulí“. Ach, angličtina, aký pôvabný jazyk.)

Keysar a Henly (2002) testovali kalibráciu rečníkov: Zvyknú rečníci podceňovať, preceňovať, alebo správne odhadovať, ako často im poslucháči rozumejú?[40] Rečníci dostali nejednoznačné vety („Muž naháňa ženu na bicykli“) a rozlišujúce obrázky (muž beží za ženou, ktorá ide na bicykli), potom ich požiadali, aby vyslovili tieto slová pred poslucháčom, a potom ich požiadali o odhad, koľko poslucháčov pochopilo zamýšľaný význam. Rečníci si mysleli, že boli správne pochopení v 72 % prípadov, ale naozaj boli pochopení v 61 % prípadov. Keď poslucháč nepochopil, rečníci si v 46 % prípadov mysleli, že pochopil; keď poslucháč pochopil, rečníci si iba v 12 % prípadov mysleli, že nepochopil.

Ďalšie pokusné osoby, ktoré *počuli* vysvetlenie, nevykazovali toto skreslenie a očakávali pochopenie poslucháčov iba v 56 % prípadov.

---

35 Boaz Keysar, „The Illusory Transparency of Intention: Linguistic Perspective Taking in Text,“ [Iluzórna priehľadnosť zámeru: Používanie jazykovednej perspektívy v texte] *Cognitive Psychology* 26 (2 1994): 165–208, doi:10.1006/cogp.1994.1006.

36 Keysar and Barr, „Self-Anchoring in Conversation.“

37 Boaz Keysar, „Language Users as Problem Solvers: Just What Ambiguity Problem Do They Solve?,“ [Používatelia jazyka ako riešitelia problémov: Aký problém nejednoznačnosti vlastne rieši?] in *Social and Cognitive Approaches to Interpersonal Communication*, ed. Susan R. Fussell and Roger J. Kreuz (Mahwah, NJ: Lawrence Erlbaum Associates, 1998), 175–200.

38 Keysar and Barr, „Self-Anchoring in Conversation.“

39 Boaz Keysar and Bridget Bly, „Intuitions of the Transparency of Idioms: Can One Keep a Secret by Spilling the Beans?,“ [Intuície o priehľadnosti slovných spojení: Môže niekto udržať tajomstvo tým, že všetko vyklopí?] *Journal of Memory and Language* 34 (1 1995): 89–109, doi:10.1006/jmla.1995.1005.

40 Boaz Keysar and Anne S. Henly, „Speakers‘ Overestimation of Their Effectiveness,“ [Ako rečníci preceňujú svoju účinnosť] *Psychological Science* 13 (3 2002): 207–212, doi:10.1111/1467-9280.00439.

Ako poznamenávajú Keysar a Barr (2002), dva dni pred útokom Nemecka na Poľsko, Chamberlain poslal list s úmyslom vysvetliť, že Británia sa v prípade invázie zapojí do boja.[41] Tento list, formulovaný zdvorilým diplomatickým jazykom, Hitler pochopil ako zmierlivý – a tanky vyštartovali.

Nebuďte príliš rýchli pri odsudzovaní tých, ktorí nepochopili vaše dokonale jasné vety, vyslovené alebo napísané. Je možné, že vaše slová boli dvojznačnejšie než si myslíte.

⁂→

9. Očakávame krátke inferenčné vzdialenosti

Životné prostredie, na ktoré je *Homo sapiens* evolučne prispôsobený→ (skrátene: „praveké prostredie“), sa skladalo z bánd→ po nanajvýš 200→ lovcov a zberačov, bez písma. Všetky zdedené vedomosti sa odovzdávali slovne, naspamäť.

V takomto svete sú všetky základné poznatky všeobecne známe. Každá informácia, ktorá nie je prísne súkromná, je verejná, bodka.

V pravekom prostredí ste sa ťažko od hocikoho vzdialili o viac ako *jeden inferenčný krok*. Keď ste objavili novú oázu, nemuseli ste svojim súkmeňovcom vysvetľovať, čo je to oáza alebo prečo je dobré piť vodu alebo ako sa chodí. Iba vy ste vedeli, kde sa tá oáza nachádza; to bola súkromná vedomosť. Ale každý mal základy na to, aby pochopil váš popis oázy, pojmy potrebné na rozmýšľanie o vode; to boli všeobecné vedomosti. Keď ste v pravekom prostredí vysvetľovali veci, takmer *nikdy* ste nemuseli vysvetľovať použité pojmy. V krajnom prípade ste museli vysvetliť *jeden* nový pojem, nie dva alebo viac naraz.

V pravekom prostredí neexistovali abstraktné vedné odbory s hromadami starostlivo zozbieraných dôkazových materiálov, zovšeobecnenými do elegantných teórií prenášaných písomne knihami, ktorých závery sú *stovky inferenčných krokov vzdialené* od všeobecne spoločných základných predpokladov.

Ak v pravekom prostredí niekto niečo povedal bez zrejmého podkladu, bol to klamár alebo hlupák. Nepomysleli ste si: „Hej, tento človek možno má dobre podložené vzdelanie, o ktorom nikto z mojej bandy nikdy nepočul,“ pretože v pravekom prostredí sa dalo spoľahnúť na to, že toto sa nestáva.

A naopak, ak ste povedali niečo očividne jasné a druhý človek to tak nevidel, potom *on* bol hlupák, alebo bol úmyselne tvrdohlavý, aby vás vytočil.

A na dôvažok, ak niekto niečo povedal bez zrejmého podkladu a *očakával*, že tomu uveríte – a správal sa rozhorčene, keď ste neuverili – to musel byť *šialenec*.

Spolu s ilúziou priehľadnosti→[8] a sebaukotvením→ to podľa mňa vysvetľuje *veľa* z povestných ťažkostí, ktoré má väčšina vedcov pri komunikácii s laickou verejnosťou – dokonca aj s vedcami z iných odborov. Keď sledujem zlyhania pri vysvetľovaní, zvyčajne vidím, že vysvetľujúci urobil *jeden* krok vzad, keď by potreboval urobiť dva alebo viac krokov vzad. Alebo že poslucháči predpokladajú, že veci by mali byť jasné na jeden krok, keď na ich vysvetlenie treba dva alebo viac krokov. Obe strany konajú, akoby očakávali veľmi krátke inferenčné vzdialenosti od všeobecných vedomostí po každú novú vedomosť.

Keď sa biológ rozpráva s fyzikom, môže zdôvodniť evolúciu povedaním, že je to „najjednoduchšie vysvetlenie“. Ale nie každý na tejto Zemi bol naočkovaný legendami z dejín vedy, od Newtona po Einsteina, ktoré dávajú slovnému spojeniu „najjednoduchšie vysvetlenie“ jeho úžasný dôraz: sú to Slová Moci, ktoré sa vyslovujú pri zrode teórií a sú vytesané na ich náhrobných kameňoch. Niekomu

41 Keysar and Barr, „Self-Anchoring in Conversation.“

→ http://lesswrong.com/lw/ke/illusion_of_transparency_why_no_one_understands/

→ http://en.wikipedia.org/wiki/Evolutionary_psychology#Environment_of_evolutionary_adaptedness

→ http://en.wikipedia.org/wiki/Band_society

→ http://en.wikipedia.org/wiki/Dunbar's_number

→8 Kapitola 8. Ilúzia priehľadnosti: Prečo vám nikto nerozumie, strana 31

→ http://lesswrong.com/lw/kf/selfanchoring/

inému môže: „Ale to je najjednoduchšie vysvetlenie!“ znieť ako zaujímavý ale sotva drvivý argument; nepripadá mu to ako až taký mocný nástroj na porozumenie firemnej politike alebo na opravu pokazeného auta. Biológ je očividne zaľúbený do svojich vlastných nápadov, príliš povýšený na to, aby sa otvoril alternatívnym vysvetleniam, ktoré znejú rovnako dôveryhodne. (Keď to znie dôveryhodne mne, malo by to znieť dôveryhodne každému príčetnému členovi mojej bandy.)

A z pohľadu biológa, on chápe, ako evolúcia po prvýkrát môže znieť trochu divne – ale keď niekto odmieta evolúciu ešte po tom, čo mu biológ vysvetlil, že to je najjednoduchšie vysvetlenie, nuž, je jasné, že nevedci sú skrátka hlupáci a nemá zmysel rozprávať sa s nimi.

Jasný argument musí vystavať inferenčnú *cestu*, začínajúc od toho, čo obecenstvo *už pozná alebo prijíma*. Ak sa nevrátite dostatočne ďaleko naspäť, rozprávate sa iba sami so sebou.

Ak v hociktorom bode niečo tvrdíte bez očividného zdôvodnenia argumentmi, ktoré ste predtým predložili, obecenstvo si skrátka pomyslí, že ste blázon.

Toto sa stane, aj keď si dovolíte *viditeľne* prikladať nejakému argumentu väčšiu váhu než si *v tej chvíli* zaslúži v očiach obecenstva. Napríklad, ak budete rozprávať, akoby ste si mysleli, že „jednoduchšie vysvetlenie“ je pádny argument pre evolúciu (čo naozaj je), namiesto iba tak trochu zaujímavej myšlienky (ako to znie niekomu, kto nebol vychovaný k úcte voči Occamovej britve).

Aha, a radšej by ste nemali nijako naznačovať, že *si myslíte*, že pracujete tucet inferenčných krokov ďalej od toho, čo obecenstvo vie, alebo že *si myslíte*, že máte špeciálne základné poznatky, ktoré oni nemajú. Obecenstvo nevie nič o evolučno-psychologickom argumente pre kognitívne skreslenie podceňovať inferenčné vzdialenosti, ktoré vedie k dopravným zápcham v komunikácii. Budú si skrátka myslieť, že sa vyvyšujete.

A ak si myslíte, že dokážete pojem „systematické podceňovanie inferenčných vzdialeností“ vysvetliť stručne, iba pár slovami, mám pre vás jednu zlú správu...

*→

10. Objektív, ktorý vidí svoje chyby

Svetlo opustí Slnko, narazí do vašich šnúrok na topánkach a odrazí sa; niektoré fotóny vojdú do zreničiek vašich očí a dopadnú na vašu sietnicu; energia fotónov vyvolá nervové impulzy; nervové impulzy sa prenesú do oblastí mozgu spracovávajúcich obraz; tam sa optická informácia spracuje a rekonštruuje na 3D model, ktorý sa rozozná ako rozviazaná šnúrka na topánke; a tak veríte, že vaše šnúrky sú rozviazané.

Toto je tajomstvo *rozumného uvedomovania* – celý tento proces previazania nie je kúzlo→35, a vy ho dokážete *pochopiť*. Dokážete *pochopiť*, ako vidíte svoje šnúrky na topánkach. Dokážete *rozmýšľať* o tom, aké druhy myšlienkových procesov budú vytvárať názory odzrkadľujúce skutočnosť, a ktoré myšlienkové procesy nebudú.

Myši dokážu vidieť, ale nedokážu pochopiť videnie. *Vy* dokážete pochopiť videnie a preto dokážete robiť veci, ktoré myši nedokážu. Chvíľku sa tomu čudujte→44, lebo to naozaj je zázrak.

Myši vidia, ale nevedia, že majú zrakovú kôru, takže nevedia korigovať optické ilúzie. Myš žije v myšlienkovom svete, ktorý obsahuje mačky, diery, syr a pasce na myši – ale nie myšacie mozgy. Ich fotoaparát nedokáže odfotiť vlastný objektív. Ale my, ako ľudia, sa dokážeme pozrieť na zdanlivo neskutočný obrázok→ a uvedomiť si, že časť toho, čo vidíme, je samotný objektív. Nemusíte vždy veriť vlastným očiam, ale musíte si uvedomiť, že *máte* oči – musíte mať oddelené myšlienkové vedierka pre mapu a pre územie, pre zmysly a pre skutočnosť. Aby ste si nemysleli, že je to všedná schopnosť, spomeňte si, aká zriedkavá je v ríši zvierat.

→ http://lesswrong.com/lw/kg/expecting_short_inferential_distances/

→35 Kapitola 35. Tajomné odpovede na tajomné otázky, strana 71

→44 Kapitola 44. „Veda“ ako zastavovač zvedavosti, strana 83

→ http://www.richrock.com/gifs/optical-illusion-wheels-circles-rotating.png

Celá myšlienka vedy je, jednoducho povedané, reflektívne uvažovanie o spoľahlivejšom procese, aby obsah vašej mysle odzrkadľoval obsah sveta. To je jedna z vecí, ktoré myši nikdy nevynájdu. Keď sa zamyslíme nad záležitosťou „robenia opakovateľných pokusov na falzifikovanie teórií“, uvidíme, *prečo* to funguje. Veda nie je oddelené magistérium[→16], vzdialené od skutočného života a porozumenia bežných smrteľníkov. Veda nie je niečo, čo platí iba vnútri laboratória[→185]. Veda samotná je pochopiteľný proces v tomto svete, ktorý koreluje mozgy so skutočnosťou.

Veda dáva *zmysel*, keď sa nad ňou zamyslíte. Myši však nevedia rozmýšľať o myslení, a preto nemajú vedu. Človek by nemal prehliadať tento zázrak – alebo potenciálnu moc, ktorú nám dáva ako jednotlivcom, nielen vedeckým spoločnostiam.

Pripúšťam, že pochopiť stroj na myšlienky môže byť *o čosi zložitejšie* než pochopiť parný stroj – ale nie je to *principiálne* odlišná úloha.

Raz som šiel na EFNet-ovský kanál #philosophy a opýtal som sa: „Veríte, že v najbližších 20 rokoch nastane jadrová vojna? Ak nie, prečo nie?“ Jeden človek mi na túto otázku odpovedal, že neočakáva jadrovú vojnu do 100 rokov, pretože „tí, ktorí môžu rozhodovať o jadrovej vojne, o ňu momentálne nemajú záujem.“ „Ale prečo to predlžuješ na 100 rokov?“ opýtal som sa. „Číra nádej,“ odpovedal.

Keď sa zamyslíme nad celým týmto myšlienkových procesom, vidíme, že myšlienka na jadrovú vojnu robí daného človeka nešťastným, a vidíme, že jeho mozog preto tento názor odmieta. Ale ak si predstavíte miliardu svetov – Everettove vetvy alebo Tegmarkove duplikáty[→,42] – tento myšlienkový proces nebude systematicky korelovať[→21] optimistov s vetvami, v ktorých jadrová vojna nenastane. (Niekto chytrý teraz určite povie: „Aha, ale pretože mám nádej, budem v práci makať o čosi viac, tým pozdvihnem globálnu ekonomiku, a tým zabránim krajinám skĺznuť do zlostného a beznádejného stavu, v ktorom je jadrová vojna možná. Takže tieto dve udalosti predsa navzájom súvisia.“ V tomto bode musíme vytiahnuť Bayesovu vetu[→] a odmerať náboj previazanosti kvantitatívne. Vaša optimistická povaha nemôže mať *taký* veľký dopad na svet; nemôže samotná znížiť pravdepodobnosť jadrovej vojny o 20% alebo o koľko vaša optimistická povaha posunula váš názor. Posunúť svoj názor o veľké množstvo kvôli udalosti, ktorá nesie iba maličký náboj previazanosti, stále pokazí vaše kreslenie mapy.)

Pýtať sa, ktorý názor vás robí šťastnými, to je pozeranie dovnútra, nie von – povie vám to niečo o vás samotných, ale nie je to indícia previazaná s prostredím. Nemám nič proti šťastiu, ale malo by vychádzať z[→2] vášho obrazu sveta, nie z manipulácie s myšlienkovými farbičkami.

Keď vidíte toto – keď vidíte, že nádej posúva vaše myšlienky *prvého rádu* neprimerane ďaleko – ak dokážete pochopiť svoj mozog ako stroj na kreslenie máp, ktorý má svoje chyby – potom môžete uplatiť reflexívnu opravu. Mozog je pokazený objektív, ktorým vidíme skutočnosť. To platí o myšacích aj ľudských mozgoch. Ale ľudský mozog je pokazený objektív, ktorý dokáže pochopiť svoje vlastné kazy – svoje systematické chyby, svoje skreslenia – a uplatniť na ne opravy druhého rádu. Toto *v praxi* robí tento pokazený objektív omnoho mocnejším. Nie dokonalým, ale omnoho mocnejším.

*→

—

→16 Kapitola 16. Nárok náboženstva na nevyvrátiteľnosť, strana 44

→185 Kapitola 185. Mimo laboratória, strana 369

→ http://arxiv.org/abs/astro-ph/0302131

42 Max Tegmark, „Parallel Universes,“ in *Science and Ultimate Reality: Quantum Theory, Cosmology, and Complexity*, ed. John D. Barrow, Paul C. W. Davies, and Charles L. Harper Jr. (New York: Cambridge University Press, 2004), 459–491.

→21 Kapitola 21. Čo je indícia?, strana 51

→ Kapitola Medzihra: Intuitívne vysvetlenie Bayesovej vety, strana 345

→2 Kapitola 2. Rozumné cítenie, strana 22

→ http://lesswrong.com/lw/jm/the_lens_that_sees_its_flaws/

B: Falošné názory

11. Nech názory platia nájomné (v očakávaných vnemoch)

Takto začína prastaré podobenstvo:

Ak v lese padne strom a nikto ho nepočuje, vznikne pritom zvuk? Jeden povie: „Áno, vznikne, pretože to spôsobí vibrácie vo vzduchu.“ Druhý povie: „Nie, nevznikne, pretože to žiaden mozog zvukovo nespracuje.“

Predstavme si, že po páde daného stromu títo dvaja pôjdu spolu do lesa. Bude jeden očakávať, že uvidí strom padnutý napravo, a druhý očakávať, že uvidí strom padnutý naľavo? Predpokladajme, že pred pádom stromu títo dvaja nechali vedľa stromu zaznamenávač zvuku. Bude jeden, keď si prehrá záznam, očakávať, že začuje niečo iné ako druhý? Predstavme si, že by všetky mozgy na svete napojili na elektroencefalograf; očakával by jeden, že uvidí iné čiary než druhý? Hoci sa títo dvaja hádajú, jeden hovorí: „Nie“ a druhý hovorí: „Áno“, neočakávajú žiadne odlišné vnemy. Títo dvaja si myslia, že majú rôzne modely sveta, ale nijako sa nelíšia v očakávaní, čo sa im *stane*.

Je lákavé pokúsiť sa túto triedu omylov odstrániť trvaním na tom, že jediným legitímnym zdrojom názoru je očakávanie zmyslového vnemu. Svet však v skutočnosti obsahuje mnohé, čo priamo nevnímame. Nevidíme atómy, z ktorých sa skladá tehla, ale tie atómy tam naozaj sú. Pod nohami máte podlahu, ale tú podlahu *nevnímate* priamo; vidíte svetlo *odrazené* od podlahy, či skôr vidíte to, čo z tohto svetla spracovala vaša sietnica a zraková kôra. Odvodiť existenciu podlahu z pohľadu na podlahu je krok naspäť do nevidených príčin vnemu. Môže to vyzerať ako veľmi krátky a priamy krok, ale stále je to krok.

Stojíte navrchu vysokej budovy, vedľa ručičkových hodín s hodinovou, minútovou a tikajúcou sekundovou ručičkou. V ruke máte kolkársku guľu a pustíte ju zo strechy. Pri ktorom tiknutí hodín budete počuť zvuk dopadu kolkárskej gule na zem?

Aby ste odpovedali presne, musíte použiť názory ako *Gravitácia Zeme je 9,8 metra za sekundu za sekundu* a *Táto budova je asi 120 metrov vysoká*. Tieto názory nie sú neverbálne očakávania zmyslového vnemu; sú to slovné výroky. Asi nie je veľkým preháňaním opísať tieto názory ako vety zložené zo slov. Avšak tieto dva názory majú inferenčný *dôsledok*, ktorý je priamym zmyslovým očakávaním – ak je sekundová ručička hodín na čísle 12, keď pustíte guľu, očakávate, že ju uvidíte na čísle 1, keď o päť sekúnd neskôr začujete úder. Aby sme predpokladali zmyslové vnemy čo najpresnejšie, musíme spracovávať názory, ktoré nie sú očakávaniami zmyslových vnemov.

Je veľkou silou druhu *Homo sapiens*, že sa dokážeme lepšie než ľubovoľný iný živočíšny druh na svete naučiť modelovať nevidené. Je to zároveň jedna z našich veľkých slabostí. Ľudia často veria na veci, ktoré sú nielen nevidené, ale aj neskutočné.

Ten istý mozog, ktorý stavia siete vydedukovaných príčin zmyslových vnemov, dokáže stavať aj siete príčin, ktoré so zmyslovými vnemami nie sú spojené alebo sú slabo spojené. Alchymisti verili, že flogiston spôsobuje oheň – mohli by sme si ich mysle veľmi zjednodušene znázorniť nakreslením malého uzla s názvom „Flogiston“ a šípky od tohto uzla k ich zmyslovému vnemu praskajúceho táboráku – lenže tento názor nedáva žiadne predpovede; linka od flogistonu k vnemu sa nastaví vždy až po danom vneme, namiesto aby ten vnem vopred ohraničila. Alebo povedzme, že vás profesor postmodernej angličtiny učí, že známy spisovateľ Wulky Wilkinsen je v skutočnosti „postutopický“. Čo by ste podľa toho mali očakávať od jeho kníh? Nič. Tento názor, ak ho tak môžeme nazvať, sa nijako nespája so zmyslovými vnemmi. Ale radšej by ste si mali zapamätať tvrdenie, že „Wulky Wilkinsen“ má vlastnosť „postutopický“, aby ste to vedeli zopakovať pri blížiacej sa písomke. Podobne, ak „postutopickí“ autori zobrazujú „koloniálne odcudzenie“; ak sa písomka bude pýtať, či Wulky Wilkinsen zobrazuje koloniálne odcudzenie, radšej by ste mali odpovedať áno. Tieto názory sú spojené jeden s druhým, hoci stále nie sú spojené so žiadnym očakávaným vnemom.

Dokážeme vybudovať celé siete názorov spojených iba navzájom – nazvime ich „vznášajúce sa“ názory. Je to jedinečná ľudská chyba medzi živočíšnymi druhmi, prekrútenie schopnosti *Homo sapiens* budovať všeobecnejšie a pružnejšie siete názorov.

Racionalistická cnosť *empirizmu* spočíva k ustavičnom pýtaní sa, aké vnemy naše názory predpovedajú – alebo ešte lepšie, zakazujú. Veríte tomu, že príčinou ohňa je flogiston? Čo potom očakávate, že sa kvôli tomu stane? Veríte, že Wulky Wilkinsen je postutopický? Čo potom očakávate, že kvôli tomu uvidíte? Nie, nie „koloniálne odcudzenie“; *aký vnem budete mať*? Veríte, že keď v lese padne strom a nikto ho nepočuje, aj tak vznikne zvuk? Aký vnem sa vám kvôli tomu musí prihodiť?

Ešte lepšie je opýtať sa: aký vnem *nemôžete* mať? Veríte, že *elan vital* vysvetľuje tajomnú živosť živých bytostí? Čo potom tento názor *nedovoľuje*, aby sa stalo – čo by definitívne vyvrátilo tento názor? Prázdna odpoveď znamená, že váš názor *neobmedzuje* vnemy; dovoľuje, aby sa vám stalo *hocičo*. Vznáša sa.

Keď sa hádate o napohľad faktickej otázke, vždy myslite na to, o akom rozdielnom očakávaní sa hádate. Ak nedokážete nájsť rozdiel v očakávaní, pravdepodobne sa hádate o označeniach vo vašej sieti názorov – alebo ešte horšie, o vznášajúcich sa názoroch, parazitoch na vašej sieti. Ak neviete, aké vnemy vyplývajú z toho, že Wulky Wilkinsen je postutopický, môžete sa hádať donekonečna. (Môžete o tom aj donekonečna publikovať články.)

Najdôležitejšie je, nepýtajte sa, čomu veriť – pýtajte sa, čo očakávať. Každá otázka o názore by mala vyplývať z otázky o očakávaní, a táto otázka o očakávaní by mala byť v centre pátrania. Každý pokus o názor by mal začať ako výsledok konkrétneho pokusu o očakávanie, a mal by naďalej platiť nájomné vo forme budúcich očakávaní. Ak sa z názoru vykľuje neplatič, vyhoďte ho.

*→

12. Bájka o vede a politike

Za Rímskej ríše sa občiansky život delil na frakcie Modrých a Zelených. Modrí a Zelení sa navzájom vraždili v súbojoch, prepadoch, skupinových bitkách, povstaniach. Procopius o týchto bojovných frakciách povedal: „Tak v nich vyrastá nenávisť voči svojim blížnym, ktorá nemá príčinu a nikdy nevyhasne ani nezmizne, pretože neustúpi ani putám manželstva, ani príbuzenstva, ani priateľstva, a je to rovnaké ešte aj keď tí, ktorí majú odlišné farby, sú bratmi či inými príbuznými.“[43] Edward Gibbon napísal: „Podpora niektorej frakcie sa stala nevyhnutnou pre každého kandidáta na občiansky alebo cirkevný post.“[44]

Kto boli títo Modrí a Zelení? Boli to športoví fanúšikovia – prívrženci modrého alebo zeleného tímu pretekárskych kočov.

Predstavte si budúcu spoločnosť, ktorá ujde do rozsiahlej siete podzemných jaskýň a zapečatí všetky vchody. Nie je dané, či utekajú pred chorobou, vojnou alebo žiarením; predpokladajme, že sa prvým Podzemšťanom podarí pestovať jedlo, nájsť vodu, recyklovať vzduch, vyrobiť svetlo a prežiť, a že ich potomkom sa darí a časom vytvoria mestá. Zo sveta hore zostali iba povesti napísané na zdrapoch papiera; jeden z takýchto zdrapov papiera opisuje *oblohu*, rozsiahly otvorený vzdušný priestor nad veľkou podlahou bez hraníc. Obloha má nebeskú farbu a obsahuje zvláštne vznášajúce sa predmety podobné ohromným balíkom bielej bavlny. Význam slova „nebeská“ je však kontroverzný; niektorí tvrdia, že označuje farbu známu ako „modrá“, iní že označuje farbu známu ako „zelená“.

V raných dobách podzemnej spoločnosti Modrí a Zelení riešili svoj spor otvoreným násilím; dnes vládne prímerie – mier zrodený z rastúceho pocitu márnosti. Zmenili sa kultúrne zvyky; je tu veľká a úspešná stredná trieda, ktorá vyrástla vďaka účinnému dodržiavaniu zákona a odvykla si od násilia. Školy poskytujú určitý historický nadhľad; ako dlho trvali bitky medzi Modrými a Zelenými, koľkí

→ http://lesswrong.com/lw/i3/making_beliefs_pay_rent_in_anticipated_experiences/

43 Procopius, *History of the Wars*, ed. Henry B. Dewing, vol. 1 (Harvard University Press, 1914).

44 Edward Gibbon, *The History of the Decline and Fall of the Roman Empire*, vol. 4 (J. & J. Harper, 1829).

zomreli, ako málo sa v dôsledku toho zmenilo. Niektorí sa otvárajú zvláštnej novej filozofii, že ľudia sú ľudia, či už Modrí alebo Zelení.

Konflikt nezmizol. Spoločnosť sa naďalej delí na Modrých a Zelených a existuje „Modrý“ a „Zelený“ postoj k takmer každej aktuálnej politicky alebo kultúrne významnej téme. Modrí presadzujú dane z osobného príjmu, Zelení presadzujú dane z obchodného obratu; Modrí presadzujú prísnejšie manželské zákony, kým Zelení chcú zjednodušiť rozvody; Modrých podporujú najmä mestské centrá, zatiaľ čo vzdialenejší farmári a predajcovia vody zvyknú byť Zelení; Modrí veria, že Zem je veľká guľatá skala v strede vesmíru, Zelení že je to plochá kamenná doska krúžiaca okolo ďalšieho objektu nazývaného Slnko. Nie každý Modrý či každý Zelený má „Modrý“ či „Zelený“ názor na všetky veci, ale sotva by ste našli mestského kupca, ktorý by veril, že obloha je modrá a zároveň by presadzoval osobné dane a voľnejšie manželské zákony.

Podzemie je stále polarizované; mier je neistý. Pár ľudí si úprimne myslí, že Modrí a Zelení by sa mali kamarátiť, a dnes už je bežné, že Zelení chodia nakupovať do Modrého obchodu, alebo že Modrí chodia do Zelenej krčmy. Z prímeria pôvodne zrodeného z vyčerpania potichu vyrastá duch tolerancie, dokonca aj priateľstva.

Jedného dňa Podzemím otrasie menšie zemetrasenie. Šesťčlennú turistickú výpravu tento otras zastihne pri prezeraní trosiek pradávnych obydlí v horných jaskyniach. Cítia krátky pohyb skaly pod nohami, jedna turistka spadne a oškrie si koleno. Skupina sa rozhodne vrátiť, v obave pred ďalšími zemetraseniami. Cestou späť jeden člen zachytí závan čohosi zvláštneho vo vzduchu, vôňu prichádzajúcu z dávno nepoužívanej chodby. Ignorujúc dobre mienené varovania svojich súputníkov, táto osoba vezme lampáš a vojde do chodby. Kamenná cesta mieri vyššie... a vyššie... až skončí v jame vykrojenej zo sveta, na mieste, kde všetky skaly končia. Diaľka, nekonečná diaľka, sa rozprestiera bez konca; námestie schopné obsiahnuť tisíce miest. V nepredstaviteľnej výške pálčivá iskra, taká jasná, že na ňu nemožno priamo pozerať, osvetľuje všetok viditeľný priestor, ako obnažené vlákno akejsi obrovskej žiarovky. Vo vzduchu visia bez podpory veľké nepochopiteľné chumáče bielej bavlny. A ten rozľahlý svietiaci strop nad tým všetkým... jeho *farba*... je...

Tu sa história rozvetvuje podľa toho, ktorý člen turistickej výpravy sa rozhodol putovať chodbou na povrch.

Aditya Modrá stála pod modrou oblohou celú večnosť a pomaly sa usmievala. Nebol to príjemný úsmev. Bola v ňom nenávisť a zranená pýcha; pripomínala si každú hádku, ktorú kedy mala so Zeleným, každé súperenie, každú prekážku v kariére. „*Mala si pravdu po celý čas,*“ šepkala jej obloha zhora, „*a teraz to môžeš dokázať.*“ Aditya tam chvíľu stála, nasávala túto správu, jasala nad ňou, a potom sa vrátila do kamennej chodby, aby to oznámila celému svetu. Ako Aditya kráčala, stískala dlaň v zaťatú päsť. „Prímerie,“ povedala, „skončilo.“

Barron Zelený dlhé sekundy nechápavo hľadel na chaotické farby. Keď prišlo pochopenie, zasiahlo ho ako úder baranidlom do žalúdka. Z očí mu vytryskli slzy. Barron si spomenul na masaker v Cathay, kde Modrá armáda zmasakrovala všetkých obyvateľov Zeleného mesta vrátane detí; spomenul si na dávneho Modrého generála Annasa Rella, ktorý označil Zelených za „jamu plnú chorôb; mor, ktorý treba vyčistiť“; pomyslel na záblesky nenávisti, ktoré vídaval v očiach Modrých, a niečo sa v ňom zlomilo. „*Ako môžeš byť na ich strane?*“ zakričal Barron na oblohu a potom začal plakať; pretože vedel, stojac pod nenávistnou modrou žiarou, že vesmír bol vždy miestom zla.

Charles Modrý zaskočene uvažoval nad modrou oblohou. Ako profesor zmiešanej univerzity Charles starostlivo zdôrazňoval, že názory Modrých aj Zelených sú rovnako platné a zaslúžia si toleranciu: Obloha je metafyzický konštrukt a nebeská je farba, ktorú možno vnímať viac ako jedným spôsobom. Charles sa nakrátko zamyslel, keby na danom mieste stál Zelený, či by nad sebou nevidel zelený strop; alebo či ten strop azda nebude zelený o takomto čase zajtra; nemohol však na to staviť ďalšie prežitie civilizácie. Toto je iba nejaký druh prírodného javu, ktorý nijako nesúvisí s morálnou filozofiou ani spoločnosťou... ľahko by sa však dal dezinterpretovať, obával sa Charles. Charles si vzdychol a vrátil sa do kamennej chodby. Zajtra sa sem vráti sám a cestu zablokuje.

Daria, kedysi Zelená, sa snažila dýchať uprostred trosiek svojho sveta. *Necúvnem*, vravela si Daria, *neodvrátim zrak.* Celý svoj život bola Zelená a teraz musí byť Modrá. Jej priatelia a jej rodina sa od nej odvrátia. *Hovor pravdu, aj keď sa ti chveje hlas*, hovorieval jej otec; ale jej otec už bol mŕtvy, a jej mama by to nikdy nepochopila. Daria opätovala oblohe pokojný modrý pohľad, pokúsila sa to prijať a nakoniec sa jej dych upokojil. *Mýlila som sa*, povedala si trúchlivo; *nie je to nakoniec také zložité.* Nájde si nových priateľov a možno jej raz aj rodina odpustí... alebo, predstavovala si s nádychom nádeje, postúpi tú istú skúšku, postaví sa pod tú istú oblohu? „Obloha je modrá," povedala Daria pokusne a nič hrozné sa nestalo; nedokázala sa však usmievať. Daria Modrá smutne vydýchla a vrátila sa do sveta, rozmýšľajúc, čo povie.

Eddin Zelený pozrel na modrú oblohu a začal sa cynicky rehotať. Dejiny jeho sveta sa konečne vyjasnili; ani on nedokázal uveriť, že mohli byť až takí blázni. „Hlúposť," povedal Eddin, „hlúposť, *hlúposť*, a po celý čas to bolo priamo tu." Nenávisť, vraždy, vojny, a po celý čas išlo len o nejakú *vec*, o ktorej niekto písal tak, ako o hocičom inom. Ani poézia, ani nič krásne, nič na čom by sa príčetnému človeku mohlo niekedy záležať, iba jedna blbosť, ktorá bola nafúknutá do nenormálnych rozmerov. Eddin sa unavene oprel o ústie jaskyne a rozmýšľal, ako tejto informácii zabráni zničiť celý svet, hoci uvažoval, či by si to vlastne všetci nezaslúžili.

Ferris mimovoľne vzdychol, zamrazený čírym úžasom a potešením. Ferris hladne pozeral okolo seba, postupne sa na všetko zrakom prisával a len neochotne odtŕhal, aby pozrel ďalej; modrá *obloha*, biele *oblaky*, šíry neznámy *priestor* plný miest a vecí (a ľudí?), ktoré žiaden Podzemšťan ešte nevidel. „Aha, tak *takúto* farbu to má," povedal Ferris a pokračoval v skúmaní.

*→

13. Viera vo vieru

Carl Sagan raz rozprával podobenstvo o človeku, ktorý k nám príde a tvrdí: „V mojej garáži je drak." Fascinujúce! Odpovedáme, že si toho draka chceme pozrieť – poďme ihneď do garáže! „Ale počkajte," povie nám dotyčný, „to je *neviditeľný* drak."

Ako Sagan podotýka, to ešte nerobí túto hypotézu nefalzifikovateľnou. Možno pôjdeme do garáže dotyčného a hoci neuvidíme žiadneho draka, budeme počuť ťažké dýchanie bez viditeľného zdroja; na podlahe sa budú tajomne zjavovať odtlačky nôh; a prístroje ukážu, že čosi v garáži spotrebováva kyslík a vydychuje kysličník uhličitý.

Predpokladajme však, že dotyčnému povieme: „Dobre, navštívime garáž a uvidíme, či budeme počuť ťažké dýchanie," a dotyčný rýchlo povie: nie, to je *nepočuteľný* drak. Navrhneme odmerať kysličník uhličitý vo vzduchu a dotyčný povie, že tento drak nedýcha. Navrhneme rozprášiť do vzduchu vrece múky a sledovať, či obklopí obrysy neviditeľného draka, a dotyčný ihneď povie: „Cez tohto draka múka prechádza."

Carl Sagan použil toto podobenstvo na vykreslenie klasického ponaučenia, že zlé hypotézy vyžadujú veľa rýchlych dodatkov, aby sa vyhli falzifikácii. Ja však toto podobenstvo hovorím kvôli inej pointe: Dotyčný musí mať *niekde* v mysli presný model tejto situácie, pretože dokáže *presne* predpovedať, *z ktorých experimentálnych výsledkov sa potrebuje vyhovoriť*.

Niektorí filozofi sú takýmito scenármi veľmi zmätení a pýtajú sa: „Verí teda dotyčný *naozaj*, že je tam drak, alebo nie?" Akoby ľudský mozog mal na disku dosť miesta iba pre jeden názor zároveň! Skutočné mysle sú omnoho zamotanejšie. Ako sme si povedali vo včerajšom článku, existujú rôzne typy názorov; nie každý názor je priame očakávanie[→11]. Dotyčný očividne *neočakáva*, že uvidí niečo nezvyčajné, keď otvorí dvere garáže; inak by sa dopredu nevyhováral. Množina výrokových názorov dotyčného môže zároveň obsahovať *V mojej garáži je drak.* Racionalistovi sa môže zdať, že tieto dva

→ http://lesswrong.com/lw/gt/a_fable_of_science_and_politics/

→11 Kapitola 11. Nech názory platia nájomné (v očakávaných vnemoch), strana 36

názory by sa mali zraziť a pobiť, aj keď sú rôzneho typu. Je však fyzikálny fakt, že môžete napísať „Obloha je zelená!“ vedľa obrázku modrej oblohy, a papier kvôli tomu nevzbĺkne plameňom.

Racionalistická cnosť empirizmu nás má chrániť pred týmto druhom chyby. Máme sa stále pýtať svojich názorov, aké vnemy predpovedajú, aby platili nájomné formou očakávania. Avšak problém dotyčného s drakom je hlbší a nedá sa vyliečiť takouto jednoduchou radou. Nie je vyslovene *ťažké* spojiť vieru v draka s očakávaným vnemom v garáži. Ak veríš, že v tvojej garáži je drak, môžeš očakávať, že otvoríš dvere a draka uvidíš. Ak draka neuvidíš, potom to znamená, že v tvojej garáži drak nie je. To je pomerne priamočiare. Môžete si to vyskúšať vo vlastnej garáži.

Nie, táto vec s neviditeľnosťou je príznakom niečoho omnoho horšieho.

Podľa toho, aké ste mali detstvo, možno si pamätáte obdobie, keď ste prvýkrát začali pochybovať o existencii Deda Mráza, ale stále ste verili, že by ste *mali* veriť v Deda Mráza, tak ste tieto pochybnosti skúšali potlačiť. Ako si všimol Daniel Dennett, kde je ťažké v niečo veriť, je často omnoho ľahšie veriť, že *by ste* tomu *mali* veriť. Čo to znamená veriť, že Najvyššia Vesmírna Obloha→12 je zároveň dokonale modrá aj dokonale zelená? Táto veta je zmätená; nie je vôbec jasné, čo by *znamenalo* veriť tomu – čomu presne *by ste* verili, keby ste verili. Omnoho ľahšie je veriť, že je *správne*, že je *dobré* a *cnostné* a *užitočné* veriť tomu, že Najvyššia Vesmírna Obloha je zároveň dokonale modrá aj dokonale zelená. Dennett toto nazýva „viera vo vieru“.[45]

A tu sa veci skomplikujú, ako to už ľudské mysle zvyknú robiť – myslím si, že ešte aj Dennett príliš zjednodušuje, ako táto psychológia funguje v praxi. Ak totiž veríte, že veríte, nemôžete sami sebe priznať, že iba veríte, že veríte, pretože cnostné je *veriť*, nie veriť, že veríte, takže ak iba veríte, že veríte, namiesto aby ste verili, nie ste cnostní. Nikto sám sebe *neprizná*: „Neverím, že Najvyššia Vesmírna Obloha je modrá a zelená, ale verím, že by som tomu mal veriť“ – jedine, že by bol nezvyčajne schopný priznať si svoj nedostatok cnosti. Ľudia neveria, že veria, že veria, oni skrátka veria, že veria.

(Ak ste z tohto zmätení, možno vám pomôže študovať matematickú logiku, kde sa vycvičíte robiť ostré rozdiely medzi výrokom P, dôkazom výroku P, a dôkazom že P je dokázateľné. Podobne ostré rozdiely sú medzi P, chcieť P, veriť v P, chcieť veriť v P, a veriť že veríte v P.)

Existujú rôzne druhy viery vo vieru. Môžete veriť vo vieru explicitne; môžete si vo svojom prúde vedomia recitovať vetu: „Je cnostné veriť, že Najvyššia Vesmírna Obloha je dokonale modrá aj dokonale zelená.“ (A pritom veriť, že tomu veríte, pokiaľ nie ste nezvyčajne schopní priznať si svoj nedostatok cnosti.) Ale sú aj menej explicitné formy viery vo vieru. Možno sa dotyčný s drakom bojí verejného výsmechu, ktorý si predstavuje ako následok verejného priznania, že sa mýlil (hoci racionalista by mu v skutočnosti zablahoželal a ostatní sa mu skôr budú vysmievať, ak bude naďalej tvrdiť, že má draka v garáži). Možno dotyčný s drakom uhýba pred predstavou priznania sebe samému, že žiaden drak nie je, pretože je to v rozpore s jeho sebaobrazom slávneho objaviteľa draka, ktorý videl vo svojej garáži to, čo sa ostatným nepodarilo vidieť.

Keby všetky naše myšlienky boli vedomé vety, s akými narábajú filozofi, ľudská myseľ by bola pre ľudí omnoho zrozumiteľnejšia. Prchavé myšlienkové obrazy, nevyslovené cúvnutia, riadenie sa nepriznanými túžbami – tie nás utvárajú tak isto, ako slová.

Aj keď s Dennettom nesúhlasím v niektorých detailoch a komplikáciách, stále si myslím, že Dennettov pojem *viery vo vieru* je kľúčovým vhľadom potrebným na pochopenie dotyčného s drakom. Potrebujeme však širší pojem *viery*, neohraničený na verbálne vety. „Viera“ by mala zahŕňať aj nevyslovené ovládače očakávania. „Viera vo vieru“ by mala zahŕňať aj nevyslovené kognitívne-behaviorálne smernice. Nie je psychologicky realistické povedať: „Dotyčný neverí, že v jeho garáži je drak; verí, že je užitočné veriť, že v jeho garáži je drak.“ Ale je realistické povedať, že dotyčný *má očakávania akoby* v jeho garáži nebol žiaden drak, ale *má výhovorky akoby* veril, že verí.

Môžete mať obyčajný myšlienkový obraz svojej garáže, bez drakov, ktorý správne predpovedá vaše vnemy po otvorení dverí, a pritom si nikdy verbálne nepomyslieť vetu *V mojej garáži nie je žiaden drak.*

→12 Kapitola 12. Bájka o vede a politike, strana 37

45 Daniel C. Dennett, *Breaking the Spell: Religion as a Natural Phenomenon* (Penguin, 2006).

Dokonca by som sa stavil, že sa vám to stalo – že keď otvárate dvere na svojej garáži alebo na spálni alebo hocikde, a nečakáte žiadnych drakov, že vám hlavou nebehá žiadna takáto verbálna veta.

A aby ste cúvli pred vzdaním sa svojej viery v draka – alebo cúvli pred vzdaním sa svojho *sebaobrazu* ako človeka, ktorý verí v draka – nie je nutné explicitne si myslieť: *Chcem veriť, že v mojej garáži je drak.* Nutné je iba cúvnuť pred predstavou priznania si, že neveríte.

Aby dotyčný s drakom správne predpovedal, z ktorých experimentálnych výsledkov sa bude potrebovať vyhovoriť, musí (a) mať niekde v mysli správny model riadiaci očakávania a (b) kognitívne jednať tak, aby chránil (b1) svoju voľne sa vznášajúcu výrokovú vieru v draka alebo (b2) svoj sebaobraz, ako verí v draka.

Ak niekto verí, že verí v draka, a zároveň verí v draka, problém je omnoho menej vážny. Taký človek bude ochotný ísť s kožou na trh ohľadom experimentálnych predpovedí a možno aj súhlasiť, že sa vzdá viery, ak sa experimentálny predpoklad ukáže nesprávny – aj keď viera vo veru môže do tohto stále zasiahnuť, ak samotná viera nie je úplne pevná. Ak si niekto chystá výhovorky *vopred*, zdá sa, že jeho viera a viera vo vieru už nie sú v súlade.

*→

14. Bayesovské džudo

Môžete sa trochu zabaviť s ľuďmi, ktorých očakávania sa rozladili s tým, čomu veria, že veria[→13].

Raz som bol na večernej párty, skúšal som jednému mužovi vysvetliť, čím sa živím, a on povedal: „Neverím, že je možná umelá inteligencia, lebo iba Boh dokáže stvoriť dušu.“

V tej chvíli som musel byť zhora osvietený, lebo som okamžite odpovedal: „Chcete tým povedať, že ak vytvorím umelú inteligenciu, vyvráti to vaše náboženstvo?“

On na to: „Čože?“

Povedal som: „No, ak vaše náboženstvo tvrdí, že je nemožné, aby som vytvoril umelú inteligenciu, potom, ak vytvorím umelú inteligenciu, bude to znamenať, že vaše náboženstvo nemá pravdu. Buď vaše náboženstvo pripúšťa, že môžem vytvoriť UI; alebo, ak zostavím UI, vyvráti to vaše náboženstvo.“

Nasledovala pauza, ako si uvedomil, že práve vyslovil hypotézu podliehajúcu falzifikácii, a potom povedal: „No, nemyslel som, že nemôžete vytvoriť inteligenciu, ale že nemôže byť emocionálna rovnakým spôsobom ako my.“

Povedal som: „Ak teda zostavím umelú inteligenciu, ktorá bez toho, že by mala naschvál naprogramovaný taký scenár, začne rozprávať o emocionálnom živote, ktorý bude znieť ako ten náš, *to* bude znamenať, že vaše náboženstvo sa mýli.“

On: „No, hm, asi sa budeme musieť zhodnúť, že sa v tomto nezhodneme.“

Ja: „Nie, to sa v skutočnosti nedá. Existuje veta o rozumnosti s názvom Aumannova veta o zhode, ktorá dokazuje, že dvaja racionalisti sa nemôžu zhodnúť na tom, že sa nezhodnú. Ak sa dvaja ľudia v niečom nezhodnú, aspoň jeden z nich musí niečo robiť nesprávne.“

Trochu sme diskutovali o tomto. Nakoniec povedal: „No, asi som tým chcel naozaj povedať, že si nemyslím, že dokážete vytvoriť niečo večné.“

Povedal som: „Nuž, to si nemyslím ani ja! Teší ma, že sme dokázali dosiahnuť zhodu, ako vyžaduje Aumannova veta o zhode.“ Podal som mu ruku, potriasol mi ju a potom odišiel.

Žena, ktorá stála obďaleč a počúvala náš rozhovor, mi vážne povedala: „Toto bolo krásne.“

„Ďakujem veľmi pekne,“ povedal som.

*→

→ http://lesswrong.com/lw/i4/belief_in_belief/

→13 Kapitola 13. Viera vo vieru, strana 39

→ http://lesswrong.com/lw/i5/bayesian_judo/

15. Predstieranie múdrosti

Najhorúcejšie miesto v pekle je vyhradené pre tých, ktorí v čase krízy zostali neutrálni.

--John F. Kennedy, nesprávne citujúc→ Danteho, známeho odborníka na peklo

Jeden príklad signalizovania dospelosti, ktorému sa oplatí venovať osobitne, je prejavovanie *neutrality* alebo *vyčkávania so záverom* s cieľom signalizovať dospelosť, múdrosť, nestrannosť alebo nadhľad.

Príkladom môže byť prípad mojich rodičov→, ktorí mi na teologické otázky ako: „Keď mal staroveký Egypt také dobré záznamy o mnohých iných veciach, prečo nemá žiaden záznam o tom, že tam niekedy boli Židia?“ odpovedali: „Ach, keď som bol v tvojom veku, tiež som kládol podobné otázky, ale už som z toho vyrástol.“

Iným príkladom môže byť riaditeľ školy, ktorý prichytí dve deti, ako sa bijú na dvore a prísne povie: „Nezáleží na tom, kto bitku začal, ale kto ju skončí.“ Samozrejme že záleží na tom, kto bitku začal. Riaditeľ zrejme nemá k dispozícii dobré *informácie* o tomto dôležitom fakte, ale ak je to tak, mal by to *povedať* a *neodmietať dôležitosť* toho, kto udrel prvý. Keby nejaký rodič skúsil udrieť riaditeľa, videli by sme, ako ďaleko by sa na súde dostal s postojom: „Nezáleží na tom, kto začal.“ Pre dospelých sú však detské bitky akurát *nepohodlné* a pre ich *pohodlie* je nepodstatné, ktoré dieťa začalo, podstatné je iba čo najrýchlejšie bitku ukončiť.

Som presvedčený, že podobne to funguje v medzinárodnej diplomacii, keď veľmoci prísne hovoria menším skupinám, aby *okamžite* prestali bojovať. Veľmociam nezáleží na tom, kto začal – kto provokoval alebo kto na provokáciu neprimerane reagoval – pretože pre veľmoci je *nepohodlné* iba pokračovanie konfliktu. Nemôžu si Izrael a Hamas skrátka podať ruky?

Toto nazývam „predstieranie Múdrosti“. Samozrejme existuje veľa spôsobov ako skúšať signalizovať múdrosť. Ale skúšať signalizovať múdrosť tým, že odmietam hádať – odmietam sčítať indície – odmietam vysloviť záver – odmietam zaujať stranu – stojím mimo sporu a pozerám zhora povýšeným a pohŕdavým pohľadom – inými slovami, signalizujem múdrosť tým, že nič nehovorím a nerobím – tak toto mi pripadá mimoriadne *pozérske*.

Paolo Freire povedal: „Umyť si ruky nad konfliktom medzi mocným a bezmocným znamená postaviť sa na stranu mocného, nie byť neutrálny.“[46] Školský dvor je skvelým miestom pre šikanujúcich a hrozným miestom pre obete, ak učiteľov nezaujíma, *kto začal*. Podobne v medzinárodnej politike: Svet, v ktorom sa veľmoci odmietajú prikloniť na niektorú stranu a iba žiadajú okamžité prímerie, je skvelý svet pre agresorov a hrozné miesto pre napadnutých. Ale samozrejme je to veľmi pohodlné miesto pre veľmoc alebo riaditeľa školy.

Časť tohto správania možno teda pripísať číremu sebectvu zo strany toho Múdreho.

Ale časť z toho súvisí aj so signalizovaním nadhľadu. Napokon – čo by si *ostatní dospelí* pomysleli o riaditeľovi, ktorý by sa naozaj *pridal* na niektorú stranu pri bitke púhych *detí*? Nuž znížilo by to jeho spoločenské postavenie na úroveň *púheho účastníka hádky*!

Podobné to má uctievaný starešina – či už je to generálny riaditeľ, prestížny akademik alebo zakladateľ diskusnej skupiny – ktorého povesť spravodlivého závisí od jeho odmietania vysloviť záver tam, kde sa ostatní pridávajú na niektorú zo strán. Strany sa dovolávajú jeho podpory, ale takmer vždy márne; pretože Múdri sú uctievaní ako sudcovia iba pod podmienkou, že takmer nikdy naozaj nesúdia – *tým* by sa z nich stali iba ďalší účastníci sporu, o nič lepší než ktorýkoľvek iný.

(Zvláštne je, že sudcovia v skutočnom systéme zákona môžu opakovane vynášať skutočné rozsudky a nestrácajú tým *automaticky* svoju povesť nestrannosti. Možno vďaka všeobecne chápanej norme, že *musia* súdiť, že to je ich náplň práce. Alebo možno preto, lebo sudcovia nemusia opakovane rozhodovať v témach rozdeľujúcich kmeň, na ktorého úcte závisia.)

→ http://www.bartleby.com/73/1211.html

→ http://lesswrong.com/lw/yo/against_maturity/

46 Paulo Freire, *The Politics of Education: Culture, Power, and Liberation* (Greenwood Publishing Group, 1985), 122.

Existujú *aj* prípady, kde je rozumné počkať s vyslovením záveru, kde ľudia rýchlo vyslovujú závery iba vďaka svojim skresleniam. Ako povedal→ Michael Rooney:

> Táto chyba sa podobá tej, ktorú stále vidím u začínajúcich študentov filozofie: keď sa stretnú s dôvodmi na skepticizmus, namiesto toho sa stanú relativistami. To znamená, že keď je rozumným postupom počkať s vyslovením záveru, priveľa ľudí namiesto toho vysloví záver, že každý záver je rovnako dôveryhodný ako hociktorý iný.

Ale ako sa teda môžeme vyhnúť (súvisiacemu, no odlišnému) pseudo-racionalistickému správaniu, keď signalizujeme svoju neskreslenú nestrannosť pomocou falošného tvrdenia, že súčasný stav indícií je vyvážený? „Veď áno, samozrejme existuje veľa zapálených darwinistov, ale ja si myslím, že nemáme dostatok dôkazov na to, aby sme sa mohli definitívne prikloniť k prirodzenému výberu namiesto inteligentného dizajnu.“

Na túto tému si odporúčam zapamätať, že aj *neutralita je definitívny záver*. Neznamená to stáť *nad* vecou. Znamená to vyjadriť definitívny a konkrétny záver, že rovnováha indícií v konkrétnej veci povoľuje *iba* jediný záver a to je neutrálny záver. Keďže aj takýto záver môže byť nesprávny, uprednostňovanie neutrality sa dá napadnúť úplne rovnako ako uprednostňovanie hociktorej strany.

Podobne v otázkach politiky. Keď niekto tvrdí, že aj strana pro-life aj strana pro-choice má dobré argumenty a že by sa v skutočnosti mali pokúsiť o kompromis a vzájomné rešpektovanie, nezastáva tým stanovisko, ktoré je *nad* oboma stranami debaty o interrupciách. Predkladá tým definitívny záver, ktorý je rovnako konkrétny ako povedanie „pro-life!“ alebo „pro-choice!“

Ak máte cieľ zlepšiť svoje všeobecné schopnosti formovať si presnejšie názory, môže byť užitočné vyhýbať sa venovaniu pozornosti emocionálne nabitým témam, ako sú potraty alebo izraelsko-palestínsky konflikt. Ale to *neznamená*, že racionalista je príliš dospelý na to, aby hovoril o politike. *Neznamená* to, že racionalista je povznesený nad tieto hlúpe škriepky, do ktorých sa zapájajú iba politickí straníci a mladiství nadšenci.

Ako opisuje Robin Hanson, schopnosť mať potenciálne rozdeľujúce konverzácie je obmedzený zdroj→. Ak si myslíte, že dokážete potiahnuť lano nabok→, máte dôvod vynaložiť svoje obmedzené zdroje na pomerne menej časté témy, kde vaša hraničná diskusia umožňuje pomerne väčší hraničný výnos.

Ale veci, ktorým ste dali menšiu prioritu, sú potom otázkou vašich obmedzených zdrojov. *Nie* otázkou vznášania sa nad vecou, pokojne a Múdro.

Moja reakcia→ na komentár Paula Grahama na Hacker News→ vyzerá ako zhrnutie, ktoré sa oplatí zopakovať.

Je rozdiel medzi:

- Vyslovením neutrálneho záveru;
- Odmietnutím investovať hraničné zdroje;
- Predstieraním, že niektoré z horeuvedeného je známkou hlbokej múdrosti, dospelosti a nadhľadu; s príslušným dôsledkom, že pôvodné strany majú menej nadhľadu a odtiaľto zhora medzi nimi nie sú podstatné rozdiely.

* →

→ http://lesswrong.com/lw/he/knowing_about_biases_can_hurt_people/dza
→ http://www.overcomingbias.com/2009/02/the-cost-of-talking-values.html
→ http://www.overcomingbias.com/2007/05/policy_tugowar.html
→ https://news.ycombinator.com/item?id=489863
→ https://news.ycombinator.com/item?id=489702
→ http://lesswrong.com/lw/yp/pretending_to_be_wise/

16. Nárok náboženstva na nevyvrátiteľnosť

Najstarší mne známy záznam vedeckého experimentu je, ironicky, príbeh o Eliášovi a Baalových kňazoch→.

Ľud Izraela kolísal medzi Jehovom a Baalom, preto Eliáš oznámil, že vykoná experiment, ktorý to rozhodne – za oných čias dosť nová myšlienka! Baalovi kňazi položia na oltár svojho býka, Eliáš položí na oltár Jehovovho býka, ale žiaden z nich nesmie zapáliť oheň; ktorý Boh je pravý, ten zošle oheň na svoju obetu. Baalovi kňazi slúžili Eliášovi ako kontrolná skupina – rovnaké drevené palivo, rovnaký býk, rovnakí kňazi, avšak vzývajúci nepravého boha. Potom Eliáš na svoj oltár nalial vodu – čím pokazil symetriu experimentu, ale to bolo ešte za dávnych čias – aby znázornil vedomé prijatie bremena dôkazu, ako keď sa požaduje hladina spoľahlivosti 0,05. Oheň prišiel zhora na Eliášov oltár, čo je experimentálne pozorovanie. Prizerajúci sa izraelský ľud kričal: „Hospodin je Boh!“ – peer review.

A potom ľudia odtiahli 450 Baalových kňazov k rieke Kišon a podrezali im hrdlá. Je to tvrdé, ale potrebné. Falzifikovanú hypotézu musíte jasne vyradiť, a to rýchlo, skôr než si začne vytvárať ochranné výhovorky. Keby Baalovým kňazom dovolili prežiť, začali by bľabotať o tom, že náboženstvo je oddelené magistérium, ktoré nemožno ani dokázať ani vyvrátiť.

Vtedy za starých čias ľudia naozaj verili→11 svojim náboženstvám namiesto toho, aby iba verili v ne→13. Biblickí archeológovia, ktorí hľadali Noemovu archu, si nemysleli, že márnia svoj čas; očakávali, že sa môžu presláviť. Až keď sa im nepodarilo nájsť potvrdzujúce indície – a namiesto nich našli odporujúce indície – urobili religionisti to, čo William Bartley nazval *útočiskom v záväzku*: „Verím, pretože verím.“

Vtedy za starých čias neexistovala predstava, že náboženstvo je oddelené magistérium. Starý Zákon je záznamom kultúrneho prúdu vedomia: história, zákon, mravné prirovnania, a takisto aj modely, ako funguje vesmír. V žiadnej pasáži Starého Zákona nenájdete nikoho hovoriť o nadprirodzenom zázraku zložitosti vesmíru. Ale nájdete kopec vedeckých tvrdení→, napríklad že vesmír bol stvorený za šesť dní (čo je metafora pre Big Bang), alebo že zajace prežúvajú svoju potravu. (Čo je metafora pre...)

Vtedy za starých čias, keby ste povedali, že miestne náboženstvo „sa nedá dokázať“, upálili by vás na hranici. Jedným zo základných kameňov viery ortodoxného judaizmu je, že Boh sa objavil na hore Sinaj a povedal hromovým hlasom: „Áno, to všetko je pravda.“ Z bayesovského pohľadu→ je to sakra jednoznačná indícia pre nadľudsky mocnú bytosť. (Hoci to samotné ešte nedokazuje, že táto bytosť je Boh, alebo že táto bytosť je dobrá – mohli to byť mimozemskí puberťáci.) Prevažná väčšina náboženstiev v ľudskej histórii – okrem tých vynájdených *veľmi* nedávno – rozpráva príbehy o udalostiach, ktoré by tvorili celkom nepochybné indície, keby sa naozaj stali. Nezávislosť náboženstva a faktických otázok je *nedávny* a výlučne *západný* prístup. Ľudia, ktorí napísali pôvodné sväté písma, tento rozdiel ani nepoznali.

Rímska ríša zdedila filozofiu od starých Grékov; zaviedla vo svojich provinciách zákon a poriadok; udržiavala úradné záznamy; a vynucovala náboženskú toleranciu. Nový Zákon, vytvorený v období Rímskej ríše, nesie v dôsledku toho stopy modernosti. Nemohli by ste si vymyslieť príbeh o tom, ako Boh úplne zničil mesto Rím (na štýl Sodomy a Gomory), pretože rímski historici by vám oponovali, a vy by ste ich nemohli jednoducho ukameňovať.

Naopak, ľudia, ktorí vymysleli príbehy Starého Zákona, si mohli vymyslieť prakticky čokoľvek, čo sa im zapáčilo. Raní egyptológovia boli úprimne šokovaní, keď nenašli absolútne žiadnu stopu po tom, že by hebrejské kmene niekedy boli v Egypte – neočakávali síce záznam o desiatich Božích ranách, ale očakávali, že nájdu aspoň *niečo*. Ako sa ukázalo, nenašli nič. Zistili, že v údajnej dobe úteku Egypt

---

→ http://www.biblia.sk/index.php?akc=biblia_sk&hl_kniha=1ki&strana=18&hl_druh=0

→11 Kapitola 11. Nech názory platia nájomné (v očakávaných vnemoch), strana 36

→13 Kapitola 13. Viera vo vieru, strana 39

→ http://www.skepticfiles.org/atheist/genesisd.htm

→ Kapitola Medzihra: Intuitívne vysvetlenie Bayesovej vety, strana 345

ovládal väčšinu Kanaánu. To je jedna *obrovská* historická chyba, ale keď neboli žiadne knižnice, nikto vám nemohol oponovať.

Rímska ríša mala knižnice. Preto Nový Zákon neobsahuje tvrdenia o veľkých, okázalých, rozľahlých geopolitických zázrakoch, ako Starý Zákon robí bežne. Namiesto toho Nový Zákon hovorí o menších zázrakoch, ktoré však stále zapadajú do rovnakej schémy indícií. Chlapec spadne a ide mu pena z úst; príčinou je nečistý duch; od nečistého ducha môžeme logicky očakávať, že ujde pred pravým prorokom, ale neujde pred šarlatánom; Ježiš vyhnal nečistého ducha; preto je Ježiš pravý prorok a nie šarlatán. Toto je úplne normálne bayesovské uvažovanie, ak vyjdeme zo základného predpokladu, že epilepsiu spôsobujú démoni (a že koniec epileptického záchvatu dokazuje, že démon ušiel).

Nielenže si náboženstvo zvyklo robiť nárok na faktické a vedecké záležitosti, náboženstvo si zvyklo robiť nárok na *všetko*. Náboženstvo predpisovalo kódex zákona – v časoch pred legislatívnymi orgánmi; náboženstvo predpisovalo históriu – v časoch pred historikmi a archeológmi; náboženstvo predpisovalo sexuálnu morálku – v časoch pred ženskou rovnoprávnosťou; náboženstvo opisovalo formy vlády – v časoch pred ústavou; a náboženstvo odpovedalo aj na vedecké otázky ohľadom biologickej taxonómie alebo vzniku hviezd. Starý Zákon nerozprával o pocite zázraku pri pohľade na zložitosť vesmíru – sústredil sa na definovanie trestu smrti pre ženy, ktoré si obliekali mužské šaty, lebo to sa v danej dobe považovalo za solídny a uspokojujúci obsah náboženstva. Moderná predstava náboženstva ako púhej *etiky* vyplýva z toho, že každú inú oblasť už prebrali vhodnejšie inštitúcie. Etika je to, čo *zostalo*.

Presnejšie povedané, ľudia *si myslia*, že etika je to, čo zostalo. Vezmite si kultúrny záznam spred 2 500 rokov. Za ten čas ľudstvo ohromne napreduje a kúsky starovekého kultúrneho záznamu sa stávajú čoraz jasnejšie zastaranými. Etika tiež nie je imúnna voči ľudskému pokroku – napríklad dnes sa pozeráme zhora na Bibliou odporúčané praktiky, ako je vlastnenie otrokov. Prečo si ľudia *myslia*, že v etike je stále dovolené hocičo?

V podstate to vôbec nie je malý etický problém, keď niekto pozabíja tisíce nevinných prvorodených chlapcov, aby tým presvedčil nevoleného faraóna, aby prepustil otrokov, ktorých sa logicky dalo z krajiny odteleportovať. Malo by to byť *jasnejšie*, než porovnateľne drobná vedecká chyba tvrdenia, že lúčne koníky majú štyri nohy. Napriek tomu, ak poviete, že Zem je plochá, ľudia na vás budú pozerať ako blázna. Ale ak poviete, že Biblia je vaším zdrojom morálky, ženy vás nezačnú fackovať. Pre väčšinu ľudí je predstava rozumnosti daná tým, čo si myslia, že im prejde; myslia si, že im prejde schvaľovanie biblickej morálky; a preto vyžaduje iba zvládnuteľné množstvo sebaklamu, aby prehliadli morálne problémy Biblie. Všetci sa dohodli na tom, že si slona v obývačke nebudú všímať a takýto stav sa dá istý čas udržať.

Možno jedného dňa ľudstvo pokročí ďalej a kto bude odporúčať Bibliu ako zdroj morálky, bude vnímaný tak, ako keď Trent Lott schvaľoval prezidentskú kampaň Stroma Thurmonda. A potom sa povie, že „skutočnou podstatou“ náboženstva bola vždy genealógia, alebo niečo.

Predstava, že náboženstvo je oddelené magistérium, ktoré *nemožno dokázať ani vyvrátiť*, je veľká lož – lož, ktorá sa opakuje znova a znova, aby ju ľudia vyslovovali bez rozmýšľania; ktorá je však, po kritickom preskúmaní, jednoducho nepravdivá. Je to hrubé skreslenie toho, ako náboženstvo historicky vzniklo, ako všetky sväté písma prezentujú svoju vieru, čo sa rozpráva deťom, aby uverili, a v čo väčšina veriacich ľudí na Zemi stále verí. Musíte obdivovať túto čiru drzosť, na rovnakej úrovni, ako že *Oceánia bola vždy vo vojne s Eastáziou*. Žalobca vytasí zakrvavenú sekeru, a obžalovaný, na chvíľku zaskočený, sa rýchlo zamyslí a povie: „Ale moju nevinu nemôžete vyvrátiť púhym dôkazom – to je oddelené magistérium!“

A keď nevyjde toto, schmatnite kúsok papiera a načmárajte si vlastnú priepustku z väzenia.

*→

—

→ http://lesswrong.com/lw/i8/religions_claim_to_be_nondisprovable/

17. Vyznávanie a fandenie

Raz som sa zúčastnil panelovej diskusie na tému: „Sú veda a náboženstvo zlučiteľné?“ Jedna zo žien v diskusii, pohanka, bez zastavenia rečnila o tom, ako verí, že Zem vznikla, keď sa v prvotnej priepasti narodila obrovská prvotná krava, ktorá lízaním priviedla na svet prvotného boha, ktorého potomkovia zabili prvotného obra a jeho telo použili na stvorenie Zeme, atď. Príbeh bol dlhý, plný podrobností a absurdnejší než predstava, že Zem leží na chrbte obrovskej korytnačky. A rozprávajúca očividne poznala vedu natoľko, aby to vedela.

Doteraz zápasím so slovami, ako opísať, čo som videl, keď táto žena hovorila. Hovorila s... pýchou? Spokojnosťou so sebou samou? S úmyselným vystatovaním?

Táto žena pokračovala v opisovaní jej mýtu stvorenia zdanlivo celú večnosť, aj keď pravdepodobne to bolo iba päť minút. Tá zvláštna pýcha/spokojnosť/vystatovanie očividne nejako súvisela s jej *vedomím*, že jej presvedčenia sú z vedeckého hľadiska škandalózne. A nebolo to preto, že by nenávidela vedu; v rámci panelovej diskusie zastávala názor, že veda a náboženstvo sú zlučiteľné. Hovorila dokonca o tom, že je celkom pochopiteľné, prečo Vikingovia hovorili o prvotnej priepasti, vzhľadom na krajinu, v akej žili – odargumentovala svoje vlastné náboženstvo! - a napriek tomu trvala na tom, že tomuto „verí“, a vyslovovala to so svojráznym uspokojením.

Nie som si istý, či Danielov Dennettov pojem „viery vo vieru[→13]“ siaha tak ďaleko, aby zahrnul aj túto udalosť. Toto bolo ešte čudnejšie. Neprednášala svoj mýtus o stvorení s fanatickou vierou niekoho, kto potrebuje ubezpečiť sám seba. Nesprávala sa tak, akoby očakávala, že presvedčí nás, obecenstvo – ani akoby našu vieru potrebovala na potvrdenie vlastnej.

Dennett, okrem pomenovania viery vo vieru, tvrdil aj, že mnohé z toho, čo nazývame „náboženské presvedčenie“, by sa naozaj malo študovať ako „náboženské vyznávanie“. Predstavte si, že mimozemský antropológ študuje skupinu študentov postmodernej angličtiny, ktorí napohľad všetci *veria*, že Wulky Wilkinsen bol postutopický autor. Správna otázka by nemala znieť: „Prečo všetci študenti veria tomuto čudnému názoru?“ ale „Prečo všetci píšu túto čudnú vetu na písomke?“ Lebo aj keď nejaká veta v podstate nemá zmysel, stále môžete vedieť, že sa od vás očakáva, že ju budete nahlas skandovať.

Myslím si, že Dennett je možno trochu príliš cynický, keď naznačuje, že náboženské vyznávanie je *iba* vyslovovanie názorov nahlas – väčšina ľudí je natoľko čestná, že keď povedia náboženskú vetu nahlas, cítia povinnosť povedať túto verbálnu vetu aj vo svojom vlastnom prúde vedomia.

Ale zdá sa, že ani pojem „náboženského vyznávania“ celkom nezahŕňa tvrdenie tejto pohanky o viere v prvotnú kravu. Keby ste mali vyznať náboženskú vieru k spokojnosti kňaza, či k spokojnosti spoluveriacich – sakra, keby ste mali uspokojiť svoj vlastný sebaobraz ako veriaceho človeka – museli by ste *predstierať*, že veríte, *omnoho presvedčivejšie*, než robila táto žena. Ako recitovala svoj príbeh o prvotnej krave, stále s tou čudnou vystatovačnou pýchou, ani sa *nesnažila* byť presvedčivá – ani sa nesnažila presvedčiť nás, že berie svoje vlastné náboženstvo vážne. Myslím si, že aj toto ma na tom tak zarazilo. Poznám ľudí, ktorí vedia, že veria smiešnym veciam, ale keď ich vyznávajú, dávajú omnoho viac úsilia do presviedčania seba samých, že svoju vieru berú vážne.

Nakoniec mi došlo, že táto žena sa nepokúša presvedčiť ani nás, ani seba. Jej recitovanie mýtu o stvorení vôbec *nebolo o* mýte o stvorení. Namiesto toho, svojím útočným päťminútovým bľabotaním *fandila pohanstvu*, ako keby držala vlajku na futbalovom zápase. Vlajka, ktorá hovorí: „MODRÍ, DO TOHO![→12]“, nie je vyjadrením faktu, ani pokusom presvedčiť; nemusí byť presvedčivá – je to pokrik.

Tá zvláštna vystatovačná pýcha... to bolo, akoby pochodovala nahá v sprievode gay pride. (Mimochodom, nemal by som žiadne námietky, keby *naozaj* pochodovala nahá v sprievode gay pride. Lesbismus nie je niečo, čo môže byť zničené pravdou[→2].) Nebol to len pokrik, ako keď niekto pochoduje, ale pohoršujúci pokrik, ako keď pochoduje nahý – verí, že ju nemožno uväzniť ani kritizovať, pretože to robí pre svoj pochod hrdosti.

→13 Kapitola 13. Viera vo vieru, strana 39

→12 Kapitola 12. Bájka o vede a politike, strana 37

→2 Kapitola 2. Rozumné cítenie, strana 22

Preto jej záležalo na tom, aby hovorila veci, ktoré boli viac než smiešne. Keby sa pokúšala, aby to vyznelo prijateľnejšie, bolo by to ako obliecť sa.

*→

18. Názor ako uniforma

Zatiaľ som opísal rozdiely medzi názorom ako ovplyvňovačom očakávania, vierou vo vieru, vyznávaním a fandením. Spomedzi týchto by sme názory ovplyvňujúce očakávania mohli nazvať „pravé názory“ a zvyšné formy „nepravé názory“. Pravý názor môže byť nesprávny alebo nerozumný, napríklad keď niekto naozaj očakáva, že modlitba vylieči jeho choré dieťa, ale o tých zvyšných formách by sa dalo povedať, že to „vlastne nie je názor“.

Ďalšou formou nepravého názoru je názor ako skupinová identifikácia – ako spôsob zapadnutia. Robin Hanson používa ako výbornú metaforu nosenie nezvyčajného oblečenia, skupinovej uniformy, ako je kňazovo rúcho alebo židovská čiapočka, takže ja to nazvem „názor ako uniforma“.

V pojmoch realistickej psychológie človeka, moslimovia, ktorí narazili lietadlom do World Trade Center, nepochybne vnímali sami seba ako hrdinov brániacich pravdu, spravodlivosť a islamský spôsob života pred odpornými cudzími netvormi, ako vo filme Deň nezávislosti. Iba veľmi neskúsený čudák, taký typ čudáka, ktorý ani netuší, ako svet vnímajú nečudáci, by toto povedal nahlas v krčme v Alabame. Toto jednoducho Američania nehovoria. Američania hovoria, že teroristi „nenávidia našu slobodu“ a že narazenie lietadlom do budovy je „zbabelý čin“. Nemôžete v jednej vete vysloviť slová „hrdinské sebaobetovanie“ a „samovražedný atentátnik“, dokonca ani za účelom presného popisu, ako nepriatelia vnímajú svet. Samotný *pojem* odvahy a altruizmu samovražedného atentátnika je nepriateľská uniforma – je to zrejmé, lebo takto to hovoria nepriatelia. Zbabelosť a sociopatia samovražedného atentátnika je americká uniforma. Žiadne úvodzovky nie sú dosť dobré, aby ste v nich mohli hovoriť, ako nepriateľ vníma svet; bolo by to ako ísť na Halloween prezlečený za nacistu.

Názor ako uniforma môže pomôcť vysvetliť, ako sa ľudia dokážu *zapáliť* pre nepravý názor. Samotná viera vo vieru alebo náboženské vyznávanie by len ťažko vytvorili skutočné, hlboké, mocné emocionálne účinky. Aspoň si to myslím; priznávam, že na toto nie som odborník. Ale zdá sa mi, že je to takto: Ľudia, ktorí prestali očakávať v súlade so svojím náboženstvom, môžu urobiť veľa preto, aby sami seba *presvedčili*, že sú zapálení, a toto zúfalstvo si možno pomýliť so zápalom. Nie je to však ten istý oheň, ktorý mali ako deti.

Na druhej strane, pre človeka je veľmi ľahké skutočne, vášnivo, celou svojou bytosťou patriť do nejakej skupiny, fandiť svojmu obľúbenému športovému tímu. (Na tomto základe spočíva podvod s „republikánmi alebo demokratmi“ a podobné falošné dilemy v iných krajinách, ale to je téma na iný článok.) Stotožniť sa s nejakým kmeňom je veľmi silná emócia. Ľudia za ňu zomierajú. A keď sa raz ľudia stotožnia s nejakým kmeňom, budú vyslovovať názory, ktoré sú uniformou tohto kmeňa, s plnou vášňou zapadnutia do tohto kmeňa.

*→

19. Signály na potlesk

Na Summite Singularity 2007 jeden z rečníkov požadoval demokratický nadnárodný vývoj UI. Pristúpil som teda k mikrofónu a povedal:

> Predstavte si, že skupina demokratických republík vytvorí konzorcium na vývoj UI, a že v celom tom procese bude veľa politikárčenia – niektoré záujmové skupiny budú mať nezvyčajne veľký vplyv, iné budú odstrčené – inými slovami, výsledok bude vyzerať tak ako produkty moderných demokracií. Ako alternatívu si predstavte, že skupina rebelských

→ http://lesswrong.com/lw/i6/professing_and_cheering/

→ http://lesswrong.com/lw/i7/belief_as_attire/

kockáčov vyvinie UI doma v pivnici, a prikáže tejto UI aby sa opýtala každého na celom svete – keď dodá mobil každému, kto ho nemá – a aby urobila čokoľvek, čo povie väčšina. Ktorá z týchto možností je podľa vás „demokratickejšia“, a cítili by ste sa pri niektorej z nich bezpečne?

Chcel som zistiť, či verí v pragmatickú primeranosť demokratického politického procesu alebo či verí v morálnu správnosť hlasovania. Rečník však odpovedal:

> Ten prvý scenár znie ako redakčný komentár v časopise *Reason* a ten druhý znie ako zápletka hollywoodskeho filmu.

Zmätene som sa opýtal:

> Aký druh demokratického procesu ste teda *mali* na mysli?

Rečník odpovedal:

> Niečo ako projekt ľudského genómu – to bol medzinárodne sponzorovaný výskumný projekt.

Opýtal som sa:

> Ako by rôzne záujmové skupiny riešili svoje konflikty v štruktúre ako bol projekt ľudského genómu?

A rečník povedal:

> Neviem.

Táto slovná výmena mi pripomenula výrok→ nejakého diktátora, ktorého sa pýtali, či má niekedy v úmysle posunúť svoj štátik smerom k demokracii:

> My veríme, že už máme demokratický systém. Niektoré prvky v ňom zatiaľ chýbajú, napríklad možnosť ľudu vyjadriť svoju vôľu.

Podstatou demokracie je konkrétny mechanizmus, ktorý rieši politické konflikty. Keby všetky skupiny uprednostňovali rovnaké pravidlá, nepotrebovali by sme demokraciu – spolupracovali by sme automaticky. Procesom uznášania môže byť priame väčšinové hlasovanie, volení zákonodarcovia, alebo hoci aj UI citlivá na názory voličov, ale musí to byť *niečo*. Čo to vôbec *znamená*, žiadať „demokratické“ riešenie, ak tým nemyslíme mechanizmus rozhodovania konfliktov?

Myslím, že to znamená, že ste povedali slovo „demokracia“ a preto sa čaká, že obecenstvo zatlieska. Nie je to ani tak výrokové tvrdenie, ako ekvivalent signálu „Potlesk“, ktorý oznamuje obecenstvu v štúdiu, kedy má tlieskať.

Tento príklad je zaujímavý iba tým, že som si pomýlil signál na potlesk s návrhom pravidiel, čoho výsledkom bol následný všeobecný trapas. Väčšina signálov na potlesk je omnoho jasnejšia a dokážete ich odhaliť jednoduchým testom opaku. Predstavte si napríklad, že niekto povie:

> Mali by sme zvážiť riziká a príležitosti UI.

Ak urobíte opak tohto tvrdenia, dostanete:

> Nemali by sme zvážiť riziká a príležitosti UI.

Keďže tento opak znie *nenormálne*, pôvodné tvrdenie je pravdepodobne normálne, z čoho vyplýva, že nevyjadruje novú informáciu. Existuje veľa legitímnych dôvodov na povedanie vety, ktorá by mimo kontextu nedávala informáciu. „Mali by sme zvážiť riziká a príležitosti UI“ môže byť úvodom k diskusnej téme; môže zdôrazniť dôležitosť konkrétneho návrhu ako ich zvažovať; môže kritizovať nevyvážený návrh. Odkázať na normálne tvrdenie môže ohraničenému racionalistovi sprostredkovať novú informáciu – súvislosť nemusela byť samozrejmá. Ale ak nenasleduje *nič* konkrétne, táto veta je pravdepodobne signálom na potlesk.

Som v pokušení urobiť raz prednášku, ktorá nebude obsahovať *nič iné* ako signály na potlesk a čakať, ako dlho obecenstvu potrvá, než sa začne smiať:

→ http://content.time.com/time/magazine/article/0,9171,954853,00.html

Som tu, aby som vám dnes navrhol, aby sme zvážili riziká a príležitosti pokročilej Umelej Inteligencie. Mali by sme sa vyhnúť rizikám a v rámci možností využiť príležitosti. Nemali by sme sa zbytočne vystavovať nebezpečenstvám, ktorým sa môžeme vyhnúť. Aby sme dosiahli tieto ciele, musíme plánovať múdro a rozumne. Nemali by sme konať v strachu a panike, ani sa poddávať technofóbii; nemali by sme sa však riadiť ani slepým nadšením. Mali by sme rešpektovať záujmy všetkých strán, ktorých sa Singularita týka. Musíme sa pokúsiť zabezpečiť, aby z pokročilej technológie vyplynuli výhody čo najväčšiemu počtu jednotlivcov, nie iba hŕstke vyvolených. Musíme sa v rámci možností pokúsiť vyhnúť násilným konfliktom s využitím týchto technológií; a musíme zabrániť jednotlivcom, ktorí by sa chceli zmocniť rozsiahlej ničivej kapacity. Musíme o týchto veciach rozmýšľať vopred, nie až dodatočne, keď už bude príliš neskoro s tým niečo urobiť...

* →

→ http://lesswrong.com/lw/jb/applause_lights/

C: Všímať si zmätok

20. Zameraj svoju neistotu

Budú výnosy dlhopisov rásť, klesať alebo zostanú rovnaké? Ak ste televízny expert a vašou prácou je vysvetliť tento výsledok dodatočne, nie je dôvod obávať sa. Bez ohľadu na to, *ktorá* z týchto troch možností sa ukáže ako pravdivá, budete vedieť vysvetliť, prečo tento výsledok dokonale sedí s vašou obľúbenou teóriou trhu. Nie je dôvod myslieť na tieto tri možnosti ako navzájom si nejako *odporujúce*, ako navzájom sa *vylučujúce*, pretože dostanete plný počet bodov za expertstvo bez ohľadu na to, ktorý výsledok nastane.

Ale moment! Čo ak ste *začínajúci* televízny expert a nemáte dosť skúseností na to, aby ste si dôveryhodné vysvetlenia vymysleli na mieste. Potrebujete si vopred prichystať poznámky na zajtrajšie vysielanie a na ich prípravu máte obmedzený čas. V tomto prípade by bolo užitočné vedieť, *ktorý* výsledok naozaj nastane – či budú výnosy dlhopisov rásť, klesať alebo zostanú rovnaké – pretože potom by ste si potrebovali pripraviť iba *jednu* množinu výhovoriek.

Žiaľ, nikto nedokáže predpovedať budúcnosť. Čo urobíte? Iste nepoužijete „pravdepodobnosti“. Všetci vieme zo školy[→], že „pravdepodobnosti“ sú číselká, ktoré sa objavujú vedľa slovných zadaní a tu nie sú žiadne číselká. Ešte horšie, *cítite* neistotu. Nespomínate si na *pocit* neistoty, keď ste manipulovali číselká pri slovných zadaniach. *Univerzitné lekcie matematiky* sú pekné čisté miesta a preto sa *matematika samotná* nemôže týkať životných situácií, ktoré nie sú pekné a čisté. Nechceli by ste nevhodne prenášať mysliace zručnosti z jedného kontextu do druhého.[→] Je jasné, že toto nie je otázka „pravdepodobností“.

Každopádne, máte iba 100 minút na prípravu výhovoriek. Nemôžete stráviť celých 100 minút na „rast“ a zároveň celých 100 minút na „pokles“ a zároveň celých 100 minút na „rovnaké“. Musíte si nejako určiť priority.

Keby ste potrebovali zdôvodniť svoje časové výdavky pred revíznou komisiou, museli by ste každej možnosti venovať rovnaký čas. Keďže tam nie sú napísané žiadne číselká, nemali by ste dokumentáciu, ktorou by ste zdôvodnili vynakladanie rôznych množstiev času. Akoby ste počuli revízorov: *A prečo ste, pán Finkledinger, strávili presne 42 minút výhovorkou číslo 3? Prečo nie 41 minút, alebo 43? Priznajte sa – nie ste objektívny! Subjektívne zvýhodňujete, čo sa vám zapáči!*

Avšak, uvedomíte si s malým záblesk om úľavy, že vás žiadna revízna komisia nebude hrešiť. To je dobre, pretože zajtra bude veľký oznam Federálneho rezervného systému a zdá sa nepravdepodobné, že by ceny dlhopisov zostali rovnaké. Nechcete minúť 33 vzácnych minút na výhovorku, o ktorej neočakávate, že ju budete potrebovať.

Vašu myseľ to unáša k výhovorkám, ktoré používate v televízii, prečo každá udalosť dôveryhodne zapadá do vašej teórie trhu. Ale čoskoro je jasné, že dôveryhodnosť vám tu nepomôže – všetky tri udalosti sú dôveryhodne vysvetliteľné. Zosúladenie s vašou obľúbenou teóriou trhu vám nepovie, ako si máte rozdeliť čas. Je neprekročiteľná priepasť medzi 100 minútami vášho času, ktoré sú ohraničené a vašou schopnosťou vysvetľovať, ako nejaký výsledok zapadá do vašej teórie, ktorá je neohraničená.

A predsa... ešte aj v tomto neistom stave mysle, zdá sa, že tieto tri udalosti *očakávate* rozdielne; *očakávate*, že niektoré výhovorky budete potrebovať viac než iné. A – toto je tá fascinujúca časť – keď pomyslíte na niečo, vďaka čomu sa zdá, že ceny dlhopisov *skôr* pôjdu hore, vtedy cítite, že budete *menej* potrebovať výhovorku pre pokles či zachovanie cien dlhopisov.

Zdá sa priam, akoby tu bol vzťah medzi tým, ako veľmi očakávate každý z týchto troch výsledkov a koľko času chcete stráviť pripravovaním každej výhovorky. Samozrejme, že sa tento vzťah nedá naozaj vyčísliť. Na prípravu svojho prejavu máte 100 minút, ale v tomto očakávaní nie je žiadne 100 na rozdelenie. (Dôjdete však na to, že *ak* nastane niektorý konkrétny prípad, vaša funkcia úžitku bude logaritmom času stráveného prípravou výhovorky.)

→ http://lesswrong.com/lw/i2/two_more_things_to_unlearn_from_school/

→ http://www.aft.org/sites/default/files/periodicals/Crit_Thinking.pdf

Napriek tomu... vaša myseľ sa stále vracia k myšlienke, že očakávanie je ohraničené, nie tak ako schopnosť vyhovárať sa, ale tak ako čas na prípravu výhovoriek. Možno by sme očakávanie mali brať ako *obmedzený zdroj*, tak ako peniaze. Vaším prvým impulzom je pokúsiť sa získať viac očakávania, ale čoskoro si uvedomíte, že aj keby ste mali viac očakávania, nebudete mať o nič viac času na prípravu výhovoriek. Nie, vašou jedinou možnosťou je *rozdeliť* váš *obmedzený zdroj* očakávania najlepšie ako viete.

Ste si dosť istí, že vás na hodinách štatistiky neučili nič podobné. Nevysvetlili vám, čo máte robiť, keď sa *cítite* tak strašne neisto. Nevysvetlili vám, čo máte robiť, keď vám nedajú žiadne číselká. Dokonca aj keby ste skúsili používať čísla, mohli by ste skúšať hocijaké čísla – nie je žiaden náznak, aký druh matematiky máte používať, ak vôbec nejaký! Možno by ste používali *dvojice* čísel, pravé a ľavé číslo... alebo ktovie čo ešte? (Na prípravu výhovoriek však máte iba 100 minút.)

Keby tak existovalo umenie, ako *zamerať svoju neistotu* – ako *natlačiť* čo najviac očakávania do toho výsledku, ktorý *naozaj nastane*!

Lenže ako by sa volalo takéto umenie? A aké pravidlá by malo?

*→

21. Čo je indícia?

> Veta „sneh je biely“ je *pravdivá* vtedy a iba vtedy, ak sneh je biely.
>
> --Alfred Tarski

> Povedať o tom, čo je, že to je, alebo o tom, čo nie je, že to nie je, to je *pravda.*
>
> *--Aristoteles, Metafyzika IV*

Ak vám tieto dva citáty nepripadajú ako dostatočná definícia „pravdy“, prečítajte si kapitolu Jednoduchá pravda. Tu budem hovoriť o „indícii“. (Tiež chcem hovoriť o názoroch ohľadom faktov, nie emócií či morálky, ako je rozlíšené v kapitole Rozumné cítenie.)

Kráčate po ceste, vaše šnúrky sa rozviažu. Krátko na to z nejakého zvláštneho dôvodu začnete *veriť*, že vaše šnúrky sú rozviazané. Svetlo opustí Slnko, dopadne na vaše šnúrky a odrazí sa; niektoré fotóny vojdú do zreničiek vašich očí a dopadnú na vašu sietnicu; energia fotónov spustí nervové impulzy; nervové ipulzy sa prenesú do oblastí mozgu spracovávajúcich obraz; a tam sa optická informácia spracuje a zrekonštruuje na 3D model, ktorý sa rozozná ako rozviazaná šnúrka. Je to postupnosť udalostí, reťaz príčin a následkov, medzi svetom a vaším mozgom, pomocou ktorej uveríte tomu, čomu veríte. Výsledkom tohto procesu je stav *mysle*, ktorý odráža stav vašich skutočných *šnúrok*.

Čo je *indícia*? Je to udalosť previazaná väzbou príčiny a následku s tým, čo chcete vedieť. Ak sú cieľom vášho prieskumu napríklad vaše šnúrky na topánkach, potom svetlo vstupujúce do vašich zreničiek je indícia previazaná s vašimi šnúrkami na topánkach. Toto by sa nemalo mýliť s technickým pojmom „previazanosť“ používaným vo fyzike – ja tu hovorím o „previazanosti“ iba v tom zmysle, že dve veci skončia v korelovaných stavoch kvôli väzbe príčiny a následku medzi nimi.

Nie každý vplyv vytvára takú „previazanosť“, akú potrebujeme pre indíciu. Nepomôže vám mať prístroj, ktorý zapípa vždy, keď doňho zadáte vyhrávajúce čísla v lotérii, pokiaľ stroj takisto zapípa *aj* keď doňho zadáte *prehrávajúce* čísla v lotérii. Svetlo odrazené od vašich topánok by nebolo užitočnou indíciou o vašich šnúrkach, keby fotóny skončili v rovnakom fyzickom stave bez ohľadu na to, či sú vaše šnúrky zaviazané alebo rozviazané.

Povedané abstraktne: Aby bola udalosť *indíciou o* cieli nášho skúmania, musí sa stať *rôzne*, spôsobom prepojeným s *rôznymi* možnými stavmi cieľa. (Povedané technicky: Musí existovať Shannonova vzájomná informácia medzi udalosťou indície a cieľom skúmania, relatívne voči vášmu súčasnému stavu neistoty ohľadom oboch z nich.)

→ http://lesswrong.com/lw/ia/focus_your_uncertainty/

Previazanie dokáže byť nákazlivé, *keď sa správne spracováva*, čo je dôvod, prečo potrebujete oči a mozog. Ak sa fotóny odrazia od vašich šnúrok na topánkach a narazia na kameň, kameň sa príliš nezmení. Kameň nebude odrážať šnúrky žiadnym užitočným spôsobom; nebude merateľne odlišný podľa toho, či boli vaše šnúrky zaviazané alebo rozviazané. To je dôvod, prečo sa kameň nepovažuje za užitočného svedka na súde. Fotografický film zachová previazanosť prichádzajúcich fotónov so šnúrkami na topánkach, preto samotná fotografia môže byť indíciou. Ak vaše oči a mozog fungujú správne, *vy* sa previažete so svojimi vlastnými šnúrkami.

To je dôvod, prečo racionalisti kladú taký silný dôraz na paradoxne vyzerajúce tvrdenie, že názor je skutočne hodnotný iba vtedy, keby bolo principiálne možné presvedčiť vás, aby ste si mysleli niečo iné. Keby vaša sietnica skončila v rovnakom stave bez ohľadu na to, aké svetlo na ňu dopadá, boli by ste slepí. Niektoré systémy názorov hovoria, ako pomerne priehľadný trik na posilnenie seba samých, že určité názory sú skutočne hodnotné iba vtedy, ak v ne veríte *bezpodmienečne* – bez ohľadu na to, čo vidíte, bez ohľadu na to, čo si myslíte. Váš mozog je povinný bez ohľadu na všetko skončiť v rovnakom stave. Preto máme frázu „slepá viera“. Ak to, čomu veríte, nezávisí na tom, čo vidíte, potom vás oslepili rovnako účinne, ako keby vám vypichli oči.

Ak vaše oči a mozog pracujú správne, vaše názory budú previazané s faktmi. *Rozumné myslenie produkuje názory, ktoré samotné sú indíciami.*

Ak váš jazyk hovorí pravdu, vaše rozumné názory, ktoré samotné sú indíciami, môžu pôsobiť ako indícia pre niekoho iného. Previazanosť sa dá prenášať reťazami príčiny a následku – a ak hovoríte a druhý počúva, aj to je príčina a následok. Keď poviete do mobilu „mám rozviazané šnúrky“, delíte sa o svoju previazanosť s kamarátom.

Preto sú rozumné názory nákazlivé, medzi úprimnými ľuďmi, ktorí si navzájom veria, že sú úprimní. A preto tvrdenie, že vaše názory *nie* sú nákazlivé – že veríte zo súkromných dôvodov, ktoré sa nedajú prenášať – je také podozrivé. Ak sú vaše názory previazané so skutočnosťou, potom by medzi úprimnými ľuďmi *mali* byť nákazlivé.

Ak vás model skutočnosti naznačuje, že výsledky vašich myšlienkových procesov by *nemali* byť pre druhých nákazlivé, potom váš model hovorí, že vaše názory samotné nie sú indíciami, čo znamená, že nie sú previazané so skutočnosťou. Mali by ste použiť reflektívnu opravu a prestať veriť.

Naozaj, pokiaľ *cítite*, na *intuitívnej* úrovni, čo toto všetko *znamená*, potom *automaticky* prestanete veriť. Pretože „môj názor nie je previazaný so skutočnosťou“ *znamená* „môj názor je nepresný“. Akonáhle prestanete veriť, že „veta ‚sneh je biely‘ je pravdivá“, mali by ste (automaticky!) prestať veriť, že „sneh je biely“, inak je niečo veľmi zle.

Poďte teda vysvetliť, prečo ten druh myšlienkových procesov, ktoré systematicky používate, produkuje myšlienky, ktoré odrážajú skutočnosť. Vysvetlite, prečo si myslíte, že ste *rozumný*. Prečo si myslíte, že použitím takých myšlienkových procesov, aké používate, myseľ skončí s názorom „sneh je biely“ vtedy a práve vtedy, keď sneh je biely. Ak neveríte, že výsledky vašich myšlienkových procesov sú previazané zo skutočnosťou, prečo veríte výsledkom vašich myšlienkových procesov? Je to predsa to isté, alebo by malo byť.

*→

22. Vedecká indícia, právna indícia, rozumná indícia

Predstavte si, že vám váš dobrý priateľ, policajný komisár, prísne dôverne povie, že hlavou zločinu vo vašom meste je Wulky Wilkinsen. Ako racionalista, mali by ste veriť tomuto tvrdeniu? Poviem to takto: ak pôjdete a začnete Wulkyho urážať, označil by som vás za somára. Keďže je obozretné správať sa akoby Wulky mal podstatne vyššiu než štandardnú pravdepodobnosť byť šéfom zločincov, výrok policajného komisára musel byť silná bayesovská indícia.

→ http://lesswrong.com/lw/jl/what_is_evidence/

Náš právny systém nezavrie Wulkyho do väzenia na základe výroku policajného komisára. Nie je to prijateľné ako *právna indícia*. Keby sme zavreli každého človeka, ktorého policajný komisár obviní, že je šéfom zločincov, možno by sme *spočiatku* chytili veľa zločineckých šéfov, plus pár ľudí, ktorých policajný komisár nemá rád. Moc má sklon korumpovať; postupne by sme chytali menej a menej skutočných zločineckých šéfov (ktorí by si dávali väčší pozor na zaistenie anonymity) a viac a viac nevinných obetí (neobmedzená moc priťahuje korupciu tak ako med priťahuje muchy).

To neznamená, že výrok policajného komisára nie je rozumná indícia. Stále má nahnutý pomer pravdepodobnosti a boli by ste somár, keby ste začali Wulkyho urážať. Ale na *spoločenskej* úrovni, pri dosahovaní spoločenských cieľov, úmyselne definujeme „právnu indíciu" tak, že zahŕňa iba isté druhy indície, ako napríklad osobné pozorovania policajného komisára v noci na 4. apríla. Každá právna indícia by ideálne mala byť rozumnom indíciou, ale nie naopak. Vyžadujeme mimoriadne silné dodatočné štandardy než rozumnú indíciu vyhlásime za „právnu indíciu".

Ako píšem túto vetu, 18. augusta 2007 o 20:33 pacifického času, mám na sebe biele ponožky. Ako racionalista, mali by ste veriť tomuto tvrdeniu? Áno. Mohol by som o tom svedčiť pred súdom? Áno. Je to *vedecký* výrok? Nie, pretože neexistuje žiaden pokus, ktorý môžete sami urobiť, aby ste si to overili. Veda sa skladá zo *zovšeobecnení*, ktoré sa týkajú mnohých konkrétnych situácií, takže môžete robiť nové pokusy, ktoré testujú toto zovšeobecnenie a tým si sami overiť, že toto zovšeobecnenie je pravdivé, bez spoliehania sa na niekoho autoritu. Veda je *verejne reprodukovateľné* poznanie ľudstva.

Tak ako súdny systém, aj veda ako spoločenský proces sa skladá z omylných ľudí. Chceme mať chránený súbor názorov, ktoré sú *mimoriadne* spoľahlivé. A chceme mať spoločenské pravidlá, ktoré podporujú tvorbu takéhoto poznania. Preto prijímame mimoriadne silné dodatočné štandardy než rozumné poznanie ustanovíme za „vedecké poznanie" a pridáme ho do chráneného súboru názorov. Mal by racionalista beriť v historickú existenciu Alexandra Veľkého? Áno. Máme hrubý obraz o starovekom Grécku, nedôveryhodný, ale lepší než maximálna entropia. Sme však závislí na autoritách ako Plutarch; nemôžeme ignorovať Plutarcha a overiť si všetko sami. Historické poznanie nie je vedecké poznanie.

Mal by racionalista veriť, že Slnko vyjde 18. septembra 2007? Áno – nie s úplnou istotou, ale je to dobrá stávka. (Pedanti: interpretujte to, že rotácia a obežná dráha Zeme zostanú zhruba rovnaké voči Slnku.) Je toto tvrdenie, keď tento článok píšem 18. augusta 2007, *vedecký* názor?

Môže sa zdať zvrátené odopierať prídavné meno „vedecký" tvrdeniam ako „18. septembra 2007 vyjde Slnko". Keby veda nevedela robiť predpovede budúcich udalostí – udalostí, ktoré sa *ešte nestali* – potom by bola zbytočná; nevedela by predpovedať žiaden pokus. Predpoveď, že Slnko vyjde, jednoznačne je *extrapoláciou* vedeckých zovšeobecnení. Zakladá sa na modeloch slnečnej sústavy, ktoré si sami môžete pokusne overiť.

Predstavte si však, že zostavujete pokus na overenie predpovede číslo 27 uznávanej teórie Q, v novom kontexte. Nemusíte mať žiaden konkrétny dôvod spochybňovať pravdivosť tohto názoru; chcete ho iba vyskúšať v novom kontexte. Zdá sa nebezpečné povedať *pred* vykonaním pokusu, že existuje „vedecký názor" na výsledok. Existuje „konvenčná predpoveď" alebo „predpoveď teórie Q". Ale ak už poznáte „vedecký názor" na výsledok, načo sa unúvať robiť pokus?

Dúfam, že začínate vidieť, prečo stotožňujem Vedu so *zovšeobecneniami*, a nie s históriou konkrétnych pokusov. Historická udalosť sa stane raz; zovšeobecnenie sa týka mnohých udalostí. História nemôžeme reprodukovať; vedecké zovšeobecnenia áno.

Je moja definícia „vedeckého poznania" *pravdivá*? To nie je dobre formulovaná otázka. Špeciálne štandardy, ktoré stanovujeme vede, sú pragmatické voľby. Nikde vo hviezdach ani v horách nie je napísané, že $p < 0,05$ musí byť štandardom pre vedecké publikácie. Mnohí dnes tvrdia, že 0,05 je príliš slabé a že by bolo *užitočné* znížiť to na 0,01 alebo 0,001.

Možno budúce generácie, na základe teórie, že veda je *verejné reprodukovateľné poznanie* ľudstva, budú ako „vedecké" označovať iba články uverejnené v časopisoch s otvoreným prístupom. Ak si pýtate peniaze za prístup k poznanie, je to časťou poznania *ľudstva*? Môžeme dôverovať nejakému záveru, ak ľudia musia platiť za to, aby ho mohli kritizovať? Je to *naozaj* veda?

Otázka: „Je to *naozaj* veda?“ je zle formulovaná. Je časopis s uzavretým prístupom za 20 000 dolárov ročne *naozaj* bayesovskou indíciou? Tak ako pri súkromnom ubezpečení policajného komisára, že Wulky je hlavou zločinu, myslím, že musíme odpovedať: „Áno.“ Ale mal by časopis s uzavretým prístupom byť kanonizovaný ako „veda“? Mali by sme ho prijať do špeciálneho chráneného súboru názorov? Osobne si myslím, že vede by lepšie poslúžilo rozhodnutie, že iba otvorené poznanie sa považuje za *verejný reprodukovateľný súbor vedomostí ľudstva.*

*→

23. Koľko indície treba?

Nedávno som definoval *indíciu* ako „udalosť previazanú väzbou príčiny a následku s tým, čo chcete vedieť“ a *previazanosť* ako „musí sa stať rôzne pre rôzne možné stavy cieľa“. Takže, koľko previazanosti – koľko indície – treba, aby podporila nejaký názor?

Začnime otázkou dosť jednoduchou na to, aby bola matematická: ako silne by ste sa potrebovali previazať s lotériou, aby ste ju vyhrali? Predpokladajme, že sa žrebuje zo sedemdesiatich loptičiek bez opakovania, a na výhru potrebujete uhádnuť šesť čísel. Potom máme 131 115 985 možných výherných kombinácií, čiže náhodne vybraný žreb má šancu vyhrať 1 / 131 115 985 (0,000 000 7 %). Aby ste vyhrali lotériu, potrebovali by ste indíciu dosť *vyberavú* na to, aby viditeľne uprednostnila jednu kombináciu pred jej 131 115 984 alternatívami.

Predpokladajme, že existujú nejaké testy, pomocou ktorých môžete pravdepodobnostne rozlíšiť medzi vyhrávajúcimi a prehrávajúcimi číslami v lotérii. Napríklad zadáte kombináciu do malej čiernej krabičky, ktorá vždy zapípa, ak je to vyhrávajúca kombinácia, ale má iba šancu 1/4 (25 %) zapípať, ak je kombinácia nesprávna. Bayesovskými slovami by sme povedali, že *pomer podmienených pravdepodobností* je 4 k 1. To znamená, že krabička má 4-krát väčšiu šancu zapípať, ak zadáme správnu kombináciu, v porovnaní so šancou zapípať, ak zadáme nesprávnu kombináciu.

To je stále celá kopa možných kombinácií. Ak napíšete 20 neprávnych kombinácií, krabička zapípa pri 5 z nich čistou náhodou (v priemere). Ak zadáte všetkých 131 115 985 možných kombinácií, krabičke síce naisto zapípa pri jednej vyhrávajúcej kombinácii, ale zapípa aj pri 32 778 996 prehrávajúcich kombináciách (v priemere).

Takže táto krabička vám neumožní vyhrať lotériu, ale je lepšia než nič. Ak použijete túto krabičku, vaše šance vyhrať stúpnu z 1 zo 131 115 985 na 1 z 32 778 997. Trochu ste pokročili smerom k nájdeniu vášho cieľa, pravdy, v rámci širokého priestoru možností.

Predpokladajme, že použijete inú čiernu krabičku, aby ste otestovali kombinácie *dvakrát, nezávisle*. Obe krabičky naisto zapípajú pre vyhrávajúci žreb. Ale šanca, že krabička zapípa pre prehrávajúcu kombináciu, je 1/4 *nezávisle* pre každú krabičku; šanca, že *obe* krabičky zapípajú pre prehrávajúcu kombináciu je teda 1/16. Môžeme povedať, že *kumulatívna* indícia dvoch nezávislých testov má pomer podmienených pravdepodobností 16:1. Počet prehrávajúcich žrebov lotérie, ktoré prejdú cez oba testy bude (v priemere) 8 194 749.

Keďže je 131 115 985 možných žrebov v lotérii, asi ste uhádli, že potrebujete indíciu, ktorej sila je okolo 131 115 985 k 1 – udalosť, alebo sériu udalostí, ktorá má 131 115 985-krát väčšiu šancu nastať pri vyhrávajúcej kombinácii než pri prehrávajúcej. Vlastne aj toto množstvo indície by stačilo iba na to, aby vám dalo šancu *pol na pol* vyhrať lotériu. Prečo? Lebo ak použijete filter takejto sily na 131 miliónov prehrávajúcich lístkov, jeden prehrávajúci lístok, v priemere, prejde týmto filtrom. Vyhrávajúci lístok tiež prejde filtrom. Zostanú vám teda dva lístky, ktoré prešli filtrom, iba jeden z nich je vyhrávajúci. Ak si môžete kúpiť iba jeden lístok, máte šancu vyhrať 50 %.

Lepší spôsob, ako si problém znázorniť: Na začiatku je 1 vyhrávajúci lístok a 131 115 984 prehrávajúcich lístkov, takže vaša šanca vyhrať je 1 : 131 115 984. Ak použijete jednu krabičku, šanca zapípania je 1 pre vyhrávajúci lístok a 0,25 pre prehrávajúci lístok. Takže 1 : 131 115 984 vynásobíme 1 :

→ http://lesswrong.com/lw/in/scientific_evidence_legal_evidence_rational/

0,25 a dostaneme 1 : 32 778 996. Pridanie ďalšej krabičky indície opäť vynásobí šance 1 : 0,25, takže šance sú teraz 1 vyhrávajúci lístok na 8 194 749 prehrávajúcich lístkov.

Je pohodlné merať indíciu v bitoch – nie ako bity na pevnom disku, ale matematické bity, ktoré sú koncepčne iné. Matematické bity sú logaritmy pravdepodobností pri základe 1/2. Napríklad ak sú štyri možné výsledky A, B, C a D, ktorých pravdepodobnosti sú 50%, 25%, 12,5% a 12,5% a ja vám poviem, že výsledok bol „D“, dal som vám tri bity informácie, pretože som vás informoval o výsledku, ktorého pravdepodobnosť bola 1/8.

Zhodou okolností je 131 115 984 o čosi viac než 2 na 27. Takže 14 krabičiek, alebo 28 bitov indície – udalosť, ktorá má 268 435 456 : 1-krát väčšiu šancu stať sa, ak je hypotéza lístka pravdivá než ak je nepravdivá – by posunula šance z 1 : 131 115 984 na 268 435 456 : 131 115 984, čo sa vykráti na 2 : 1. Šanca 2 ku 1 znamená dve šance na výhru pre každú šancu na prehru, čiže *pravdepodobnosť* výhry s 28 bitmi indície je 2/3. Pridanie ďalšej krabičky, ďalších 2 bitov indície, by posunulo šance na 8 : 1. Pridanie ešte ďalších dvoch krabičiek by posunulo šance na výhru na 128 : 1.

Ak teda chcete licenciu na *silné presvedčenie*, že vyhráte lotériu – ktoré tu svojvoľne definujeme ako pravdepodobnosť omylu menšiu než 1% – 34 bitov indície ohľadom vyhrávajúcej kombinácie by malo stačiť.

Vo všeobecnosti sa pravidlo na zvažovanie „koľko indície treba“ riadi podobným vzorom: Čím väčší je *priestor možností*, v ktorom leží daná hypotéza, alebo čím menej pravdepodobná táto hypotéza vyzerá *a priori* v porovnaní s jej susedmi, alebo čím viac si chcete byť istí, tým viac indície potrebujete.

Nemôžete vzdorovať týmto pravidlám; nemôžete si vytvoriť presné názory na základe nedostatočnej indície. Povedzme, že máte 10 krabičiek v rade a začnete zadávať kombinácie do krabičiek. Nemôžete sa zastaviť pri prvej kombinácii, ktorá dostane zapípanie od všetkých 10 krabičiek a povedať: „Ale veď šanca, že by sa toto stalo pre prehrávajúcu kombináciu je milión ku jednej! Môžem ignorovať tieto teoretické bayesiánske pravidlá a skončiť tu.“ V priemere, na jeden vyhrávajúci lístok prejde takýmto testom aj 131 prehrávajúcich. Ak vezmeme do úvahy priestor pravdepodobností a prvotnú nepravdepodobnosť, skočili ste k príliš silnému záveru na základe nedostatočnej indície. To nie je samoúčelná byrokratická regulácia, to je matematika.

Samozrejme, aj tak môžete *veriť* na základe nedostatočnej indície, ak je to váš rozmar; ale nebudete môcť veriť *presne*. Je to ako keby ste sa pokúšali šoférovať auto bez paliva, pretože neveríte na nejaké hala-bala predstavy, že na jazdenie treba palivo. Nebolo by omnoho *zábavnejšie* a lacnejšie, keby sme sa rozhodli zrušiť zákon, že autá potrebujú palivo? Nebolo by to očividne lepšie pre každého? Nuž, môžete skúsiť, ak je to váš rozmar. Môžete dokonca aj zavrieť oči a predstierať, že auto sa hýbe. Ale aby ste *naozaj* docestovali k presným názorom, na to potrebujete palivo indícií, a čím ďalej chcete ísť, tým viac paliva potrebujete.

* →

24. Einsteinova drzosť

V roku 1919 Sir Arthur Eddington viedol expedície do Brazílie a na ostrov Principe s cieľom pozorovať zatmenie slnka a otestovať tým experimentálnu predpoveď Einsteinovej novej teórie Všeobecnej Relativity. Novinár sa opýtal Einsteina, čo by urobil, keby Eddingtonove pozorovania neboli v súlade s jeho teóriou. Einstein sa preslávil odpoveďou: „Bolo by mi ho ľúto. Teória je správna.“

Vyzerá to ako veľmi zadubený výrok, vzdorujúci klišé Tradičnej Rozumnosti, že experiment je suverénom nad všetkým. Zdá sa, že Einstein mal takú veľkú drzosť, že by odmietol skloniť hlavu a podrobiť sa odpovedi Prírody, ako vedci musia. Kto môže *vedieť*, že jeho teória je správna, ešte pred experimentálnym testom?

Samozrejme sa ukázalo, že Einstein mal pravdu. Snažím sa nekritizovať ľudí, keď majú pravdu. Ak si naozaj kritiku zaslúžia, nebudem musieť dlho čakať, kým sa pomýlia.

→ http://lesswrong.com/lw/jn/how_much_evidence_does_it_take/

A možno Einstein nebol až taký zadubený, ako znel...

Aby ste priradili pravdepodobnosť vyššiu než 50 % správnemu kandidátovi zo súboru 100 000 000 možných hypotéz, potrebujete aspoň 27 bitov indície (plus-mínus). Nemôžete očakávať, že nájdete správneho kandidáta bez takýchto silných testov, pretože pri slabších testoch viac než jeden kandidát prejde všetkými testmi. Ak sa pokúsite použiť test, ktorý má šancu na falošný pozitívny výsledok iba milión k jednej (cca 20 bitov), zostane vám sto kandidátov. Už samotné *nájdenie* správnej odpovede v širokom priestore možností si vyžaduje veľké množstvo indície.

Tradičná Rozumnosť kladie dôraz na dokazovanie: „Ak ma chceš presvedčiť o X, musíš mi dať aspoň Y bitov indície.“ Často skĺznem do takéhoto vyjadrovania, keď poviem veci ako: „Aby sme *zdôvodnili* vieru v tento výrok s pravdepodobnosťou väčšou než 99 %, potrebujeme 34 bitov indície.“ Alebo: „aby ste vašej hypotéze priradili pravdepodobnosť vyššiu ako 50 %, potrebujete 27 bitov indície.“ Tradičné vyjadrovanie naznačuje, že začínate s nejakým tušením alebo nejakým súkromným spôsobom uvažovania, ktoré vás privedie k predkladanej hypotéze, a potom musíte zhromaždiť „indície“, aby ste to *potvrdili* – aby ste presvedčili vedeckú komunitu alebo zdôvodnili tvrdenie, že *veríte* vo svoje tušenie.

Ale z bayesiánskeho pohľadu potrebujete množstvo indície približne rovné zložitosti hypotézy už len na to, aby ste túto hypotézu objavili v priestore teórií. Nie je to otázka zdôvodňovania niečoho pred niekým. Ak existuje sto miliónov alternatív, potrebujete aspoň 27 bitov indície, aby ste vôbec zamerali svoju pozornosť výhradne na správnu odpoveď.

To je pravda, dokonca aj keď svoj odhad nazývate „tušenie“ alebo „intuícia“. Tušenia a intuície sú skutočné procesy v skutočnom mozgu. Ak váš mozog nedostane na strávenie aspoň 10 bitov skutočne previazanej platnej bayesiánskej indície, potom vám váš mozog nedokáže vyhrať správnu 10-bitovú hypotézu do pozornosti – vedome, nevedome, hocijako. Nevedomé procesy nemôžu nájsť jeden cieľ z miliónov použitím púhych 19 bitov previazanosti o nič viac než vedomé procesy. Tušenia môžu byť záhadou pre toho, kto tuší, ale nemôžu porušovať fyzikálne zákony.

Vidíte, kam toto smeruje: *Vo chvíli, keď Einstein prvýkrát sformuloval svoju hypotézu* – keď mu prvýkrát v hlave naskočili rovnice – musel *už mať* dostatočné indície z pozorovania, aby vybrali zložité rovnice Všeobecnej Relativity do jeho pozornosti. Inak by ich nemohol mať *správne*.

A teraz, aká je šanca, že by Einstein mal z pozorovania *presne* toľko indície, aby priviedla Všeobecnú Relativitu do jeho pozornosti, ale zdôvodňovala by iba pravdepodobnosť 55 %? Povedzme, že Všeobecná Relativita je 29,3-bitová hypotéza. Aká je šanca, že Einstein naďabil na *presne* 29,5 bitu indície počas svojho štúdia fyziky?

Nie veľká! Ak mal Einstein z pozorovania dosť indície, aby vôbec vybral správne rovnice Všeobecnej Relativity, potom mal pravdepodobne dosť indície, aby si bol *sakra istý*, že Všeobecná Relativita je pravdivá.

V skutočnosti, keďže ľudský mozog nevie dokonale efektívne spracovávať informácie, Einstein mal pravdepodobne *omnoho viac indície*, než by v princípe bolo treba, aby dokonalý bayesiánec priradil Všeobecnej Relativite obrovskú dôveryhodnosť.

„Bolo by mi ho ľúto; teória je správna“ neznie zďaleka tak pohoršujúco, keď sa na to pozriete z tohto pohľadu. A pamätajte, že Všeobecná Relativita *bola* správna, z celého toho rozsiahleho priestoru možností.

*→

25. Occamova britva

Čím zložitejšie je vysvetlenie, tým viac indície potrebujete už len aby ste ho našli v priestore názorov. (V Tradičnej Rozumnosti sa to často formuluje zavádzajúco ako „Čím zložitejšie je tvrdenie, tým viac indície potrebujete, aby ste zaň argumentovali.“) Ako môžeme merať zložitosť vysvetlenia? Ako môžeme určiť, koľko indície potrebujeme?

→ http://lesswrong.com/lw/jo/einsteins_arrogance/

Occamova britva sa často formuluje ako: „Najjednoduchšie vysvetlenie, ktoré je v súlade s faktmi.“ Robert Heinlein odpovedal, že najjednoduchšie vysvetlenie je: „Pani z dolného konca ulice je bosorka; ona to urobila.“

Vidno, že dĺžka anglickej vety nie je dobrým spôsobom, ako merať „zložitosť“. A „súlad“ s faktmi, daný iba tým, že ich teória *nezakazuje*, tiež nestačí.

Prečo presne je dĺžka anglickej vety zlým meradlom zložitosti? Pretože keď vyslovíte nejakú vetu, používate *označenia* pre pojmy, ktorým poslucháč rozumie – ten poslucháč má už zložitosť uloženú v sebe. Prestavme si, že celú tú Heinleinovu vetu skrátime na „Pzdkujbotu!“, takže sa celé vysvetlenie dá vyjadriť jediným slovom; alebo ešte lepšie, dajme mu ľubovoľnú krátku nálepku, napríklad: „Fnord!“ Zredukovala sa tým zložitosť? Nie, pretože musíte poslucháčovi dopredu vysvetliť, že „Pzdkujbotu!“ je skratka z: „Pani z dolného konca ulice je bosorka; ona to urobila.“ A samotné slovo „bosorka“ je nálepka pre isté mimoriadne tvrdenia – to, že ho už všetci poznáme, ešte neznamená, že samotný pojem je jednoduchý.

Obrovský elektrický blesk príde z oblohy, niečo trafí a severskí domorodci povedia: „Možno sa nejaký naozaj mocný činiteľ nahneval a hodil blesk.“ Ľudský mozog je najzložitejším artefaktom v známom vesmíre. Ak nám *hnev* pripadá jednoduchý, je to preto, lebo nevidíme všetky tie nervové obvody, ktoré túto emóciu implementujú. (Pokúste sa vysvetliť, prečo je *Saturday Night Live* zábavné, mimozemšťanovi bez zmyslu pre humor. Necíťte sa však nadradení; vy zase nemáte zmysel pre fnord.) Ľudia, ktorí vymysleli hypotézu Thora, konateľa hromu, si neuvedomovali zložitosť hnevu, a vôbec zložitosť inteligencie.

Vysvetliť *človeku* Maxwellove rovnice trvá dlhšie než vysvetliť mu Thora. Ľudia nemajú zabudovaný slovník pre integrály tak, ako máme zabudovaný slovník pre hnev. Musíte vysvetliť svoj jazyk, a jazyk za týmto jazykom, a samotný pojem matematiky, než začnete hovoriť o elektrine.

A predsa sa zdá, že by mal existovať nejaký zmysel, v ktorom sú Maxwellove rovnice *jednoduchšie* než ľudský mozog alebo Thor, konateľ hromu.

Existuje: Je *omnoho* ľahšie (ako sa ukazuje) napísať počítačový program, ktorý simuluje Maxwellove rovnice, v porovnaní s počítačovým programom, ktorý simuluje inteligentnú emocionálnu myseľ ako je Thor.

Formalizmus Solomonoffovej indukcie meria „zložitosť opisu“ dĺžkou najkratšieho počítačového programu, ktorý vypíše tento opis ako svoj výstup. Ak hovoríte o „najkratšom počítačovom programe“, ktorý niečo urobí, musíte upresniť priestor počítačových programov, čo si vyžaduje jazyk a interpreter. Solomonoffova indukcia používa Turingove stroje, alebo skôr bitové reťazce, ktoré definujú Turingove stroje. Čo ak sa vám Turingove stroje nepáčia? V tom prípade máte iba konštantnú penaltu zložitosti za vytvorenie svojho univerzálneho Turingovho stroja, ktorý intepretuje taký kód, aký mu zadáte, v tom programovacom jazyku, ktorý sa vám páči. Rôzne induktívne formalizmy sú relatívne voči sebe penalizované v najhoršom prípade konštantnou veličinou zodpovedajúcou veľkosti univerzálneho interpretera daného formalizmu.

V lepších (podľa mňa) verziách Solomonoffovej indukcie počítačový program nevytvára deterministickú predpoveď, ale priraďuje pravdepodobnosti reťazcom. Napríklad by sme mohli napísať program na vysvetlenie vyváženej mince tak, že by sme napísali program, ktorý priradí rovnakú pravdepodobnosť všetkým 2^N reťazcom dĺžky N. Toto je prístup Solomonoffovej indukcie k *súladu* s pozorovanými údajmi. Čím vyššiu pravdepodobnosť program priradí pozorovaným údajom, tým viac je program *v súlade* s údajmi. A súčet pravdepodobností musí byť 1, takže program, ktorý je v lepšom „súlade“ s jednou možnosťou, musí uberať masu pravdepodobnosti z nejakej inej možnosti, s ktorou je potom v horšom „súlade“. Neexistuje žiadna supervyvážená minca, ktorá priradí pravdepodobnosť 100 % hlave a pravdepodobnosť 100 % znaku.

Ako vyvážime súlad s dátami voči zložitosti programu? Keby sme ignorovali penaltu zložitosti a mysleli *iba* na súlad, potom by sme vždy dávali prednosť programom, ktoré tvrdia, že deterministicky predpovedajú dané údaje a priraďujú im pravdepodobnosť 100 %. Ak na minci padne „HTTHHT“, potom

program, ktorý tvrdí, že na minci vždy padá „HTTHHT“, je s pozorovanými údajmi v 64-krát väčšom súlade než program, ktorý tvrdí, že minca je vyvážená. Naopak, keby sme ignorovali súlad a brali do úvahy *iba* zložitosť, potom by hypotéza „vyváženej mince“ vyzerala vždy jednoduchšia než ľubovoľná iná hypotéza. Aj keby minca dopadla „HTHHTHHHTHHHHTHHHHHT...“ Veru, hypotéza vyváženej mince *je* jednoduchšia a je v súlade s týmito dátami rovnako dobre ako s hocijakým iným reťazcom 20 hodov mince – ani viac, ani menej – ale vidíme, že iná hypotéza, nie omnoho zložitejšia, je s danými údajmi v omnoho lepšom súlade.

Ak dovolíte programu uložiť o jeden bit informácie viac, môže tým rozdeliť priestor možností napoly a tým prideliť dvakrát väčšiu pravdepodobnosť všetkým bodom v zostávajúcom priestore. To naznačuje, že jeden bit zložitosti programu by mal stáť *aspoň* „dvakrát lepší“ súlad. Ak skúsite vytvoriť počítačový program, ktorý explicitne obsahuje výsledok „HTTHHT“, tých šesť bitov, ktoré ste stratili na zložitosti, musí zrušiť všetku dôveryhodnosť získanú 64-násobne lepším súladom. V opačnom prípade by ste skôr či neskôr došli k záveru, že všetky vyvážené mince sú vopred stanovené.

Pokiaľ váš program nie je chytrý a *nekomprimuje* údaje, nemalo by vám pomôcť presunúť jeden bit z údajov do popisu programu.

Solomonoffova indukcia predpovedá postupnosti tak, že urobíte súčet cez všetky povolené počítačové programy – ak je povolený ľubovoľný program, Solomonoffova indukcia je nevypočítateľná – kde každý program dostane prvotnú pravdepodobnosť 1/2 umocnenú na dĺžku jeho kódu v bitoch a každý program je ďalej vážený podľa svojho súladu s doteraz pozorovanými údajmi. To vám dáva váženú zmes odborníkov na predpovedanie budúcich bitov.

Formalizmus minimálnej dĺžky správy je takmer ekvivalentný Solomonoffovej indukcii. Pošlete reťazec popisujúci kód a potom pošlete reťazec popisujúci údaje v tomto kóde. Ktorékoľvek vysvetlenie vedie k najkratšej *celkovej dĺžke*, je najlepšie. Ak si predstavíte množinu povolených kódov ako priestor počítačových programov a jazyk na popis kódov ako univerzálny stroj, potom je minimálna dĺžka správy takmer ekvivalentná Solomonoffovej indukcii. (Takmer, pretože si vyberá *najkratší* program namiesto sumy cez všetky programy.)

To nám umožňuje jasne vidieť problém použitia: „Pani z dolného konca ulice je bosorka; ona to urobila“ na vysvetlenie vzoru v postupnosti „0101010101“. Ak posielate správu kamarátovi a pokúšate sa vysvetliť pozorovanú postupnosť, museli by ste povedať: „Pani z dolného konca ulice je bosorka; ona to urobila, že postupnosť vyšla 0101010101.“ Vaše obvinenie z bosoráctva by vám neumožnilo *skrátiť* zvyšok správy; stále by ste museli opísať, do najmenšieho detailu, údaje, ktoré jej bosoráctvo spôsobilo.

Bosoráctvo môže byť v súlade s našimi pozorovaniami v tom zmysle, že ich kvalitatívne *povoľuje*; ale to len preto, lebo bosoráctvo povoľuje *všetko*, rovnako ako povedanie „Flogiston!“ Takže aj keď poviete „bosorka“, stále musíte popísať všetky pozorované údaje do najmenšieho detailu. *Nezmenšili ste celkovú dĺžku správy opisujúcej vaše pozorovania* tým, že ste preniesli správu o bosoráctve; jednoducho ste len pridali zbytočný úvod, čím ste zvýšili celkovú dĺžku.

Tá skutočná zákernosť bola skrytá v slove „to“ v časti „urobila to bosorka“. *Čo* presne urobila bosorka?

Samozrejme, vďaka skresleniu spätného pohľadu[→] a ukotvovaniu a falošným vysvetleniam a falošnej kauzalite a pozitívnemu skresleniu a motivovanému poznávaniu sa môže sa zdať úplne jasné, že keď je žena bosorka, *samozrejme* spôsobí, že na minci padne 0101010101. Ale k tomu sa čoskoro dostaneme...

*→

→ http://lesswrong.com/lw/il/hindsight_bias/

→ http://lesswrong.com/lw/jp/occams_razor/

26. Tvoja sila racionalistu

Nasledujúca vec sa mi stala v diskusnej miestnosti IRC kedysi dávno, keď som sa ešte potĺkal po diskusných miestnostiach IRC. Čas mi zahmlil pamäť a môj popis môže byť nepresný.

Bol som v diskusnej miestnosti IRC, keď niekto napísal, že jeho známy potrebuje lekársku radu. Jeho známy hovorí, že mal náhle bolesti v hrudníku, tak zavolal sanitku, sanitka prišla, ale zdravotníci mu povedali, že to nič nie je a odišli, a bolesti v hrudníku sa stále zhoršujú. Čo má jeho známy robiť?

Bol som touto historkou zmätený. Pamätal som si, ako som čítal o bezdomovcoch v New Yorku, ktorí si volali sanitky, len aby ich odviezli niekam do tepla, a že ich zdravotníci vždy museli vziať na pohotovosť, hoci aj 27-krát. Pretože keby ich nevzali, majiteľa sanitiek by bolo možné žalovať o veľa a veľa peňazí. Podobne, pohotovosti majú zo zákona povinnosť poskytnúť zdravotnú starostlivosť každému bez ohľadu na jeho schopnosť platiť. (Náklady potom znáša nemocnica a sú to obrovské sumy, takže nemocnice rušia svoje pohotovosti... Človek sa čuduje, načo vôbec máme ekonómov, keď ich jednoducho ignorujeme.) Takže som celkom nerozumel, ako sa popisované udalosti mohli stať. *Každého*, kto by nahlásil náhle bolesti v hrudníku, by okamžite odvliekli na pohotovosť.

A tu som ako racionalista sklamal. Spomenul som si na pár prípadov, keď môj doktor úplne odmietol panikáriť pri popise príznakov, ktoré sa mne zdali veľmi hrozivé. A lekársky establishment mal vždy pravdu. Každý jeden raz. Raz aj mňa bolel hrudník a lekár mi trpezlivo vysvetlil, že mu opisujem bolesť hrudného svalu, nie srdcový infarkt. Povedal som teda na IRC: „Pozri, ak zdravotníci povedali tvojmu známemu, že to nič nie je, *naozaj* to nič nie je – keby bola najmenšia šanca vážneho problému, boli by ho zobrali."

Takto sa mi podarilo príbeh vysvetliť v rámci môjho existujúceho modelu, hoci som cítil, že to bolo trochu nasilu...

Neskôr sa dotyčný vrátil do diskusnej miestnosti IRC a povedal, že jeho známy si to celé vymyslel. Zrejme to nebol jeden z jeho spoľahlivejších známych.

Azda som si mohol uvedomiť, že neznámy známy známeho z IRC môže byť menej dôveryhodný→ než článok uverejnený v odbornom časopise. Žiaľ, veriť je ľahšie než neveriť; veríme inštinktívne, neveriť však vyžaduje vedomé úsilie.[47]

Namiesto toho som silným tlakom donútil svoj model skutočnosti, aby vysvetlil nezrovnalosť, ktorá sa *v skutočnosti nikdy nestala*. A *vedel som*, aké je to trápne. *Vedel som*, že užitočnosť modelu nie je v tom, čo dokáže vysvetliť, ale v tom, čo nedokáže. Hypotéza, ktorá nič nezakazuje, povoľuje všetko a preto nedokáže obmedziť očakávanie.

Tvoja sila racionalistu je tvoja schopnosť byť viac zmätený výmyslom než skutočnosťou. Ak dokážeš rovnako dobre vysvetliť hocijaký výsledok, máš nulové vedomosti.

Všetci sme z času na čas slabí; smutné je, že som *mohol* byť silnejší. Mal som všetky informácie potrebné na odvodenie správnej odpovede, dokonca som si *všimol* problém, ale potom som ho ignoroval. Môj pocit zmätenia bola Nápoveda a ja som túto Nápovedu zahodil.

Mal som venovať viac pozornosti tomu pocitu, že *to bolo trochu nasilu*. Je to jeden z najdôležitejších pocitov, aké môže mať hľadač pravdy, časť tvojej sily racionalistu. Je to chybou dizajnu ľudského rozumu, že sa tento pocit prejavuje ako tichý hlások v pozadí mysle namiesto kvíliacej poplašnej sirény a žiarivého neónového nápisu:

BUĎ JE TVOJ MODEL NESPRÁVNY, ALEBO TENTO PRÍBEH NIE JE PRAVDA.

*→

→ http://www.overcomingbias.com/2007/08/truth-bias.html

47 Daniel T. Gilbert, Romin W. Tafarodi, and Patrick S. Malone, „You Can't Not Believe Everything You Read," *Journal of Personality and Social Psychology* 65 (2 1993): 221–233, doi:10.1037/0022-3514.65.2.221.

→ http://lesswrong.com/lw/if/your_strength_as_a_rationalist/

27. Neprítomnosť indície je indíciou neprítomnosti

Z knihy Robyna Dawesa *Rozumná voľba v neistom svete*:[48]

> Post-hoc napasovanie indícií do hypotézy bolo súčasťou najsmutnejšej kapitoly dejín Spojených Štátov: uväznenia Američanov japonského pôvodu v internačných táboroch začiatkom druhej svetovej vojny. Keď kalifornský guvernér Earl Warren vypovedal 21. februára 1942 v San Franciscu pred komisiou kongresu, v rámci kladenia otázok mu pripomenuli, že do tej doby sa nevyskytla žiadna sabotáž ani iný typ špionáže, ktorú by mali na svedomí Američania japonského pôvodu. Warren odpovedal: „Som toho názoru, že táto neprítomnosť [podvratnej aktivity] je tým najhrozivejším znamením celej našej situácie. Presviedča ma to azda viac než hociktorý iný faktor, že sabotáže, ktoré prídu, že aktivity piatej kolóny, ktoré prídu, sú načasované rovnako ako bol načasovaný Pearl Harbor... Verím tomu, že sa nás snažia učičíkať do falošného pocitu bezpečia."

Vezmime si tento Warrenov argument z bayesovského hľadiska. Keď vidíme indíciu, hypotézy, ktoré tejto indícii pripisovali *vyššiu* podmienenú pravdepodobnosť, získajú pravdepodobnosť na úkor hypotéz, ktoré tejto indícii pripisovali *nižšiu* podmienenú pravdepodobnosť. Toto je jav *relatívnych* podmienených pravdepodobností a *relatívnych* pravdepodobností. Môžete danej indícii pripísať vysokú podmienenú pravdepodobnosť a predsa stratiť pravdepodobnosť v prospech nejakej inej hypotézy, pokiaľ táto iná hypotéza pripísala ešte vyššou podmienenú pravdepodobnosť.

Warren sa snaží argumentovať, že neprítomnosť sabotáže *potvrdzuje*, že piata kolóna existuje. Môžeme argumentovať, že piata kolóna *možno* svoje sabotáže odkladá na neskôr. Ale aj tak je vyššia podmienená pravdepodobnosť, že *neexistencia* piatej kolóny spôsobuje neprítomnosť sabotáže.

Nech E je pozorovanie sabotáže, a ~E pozorovanie neprítomnosti samotáže. Symbol H1 označuje hypotézu piatej kolóny Američanov japonského pôvodu, a H2 hypotézu, že takáto piata kolóna neexistuje. *Podmienená pravdepodobnosť* P(E|H), alebo „E za predpokladu H" je s akou istotou očakávame, že uvidíme indíciu E, ak predpokladáme, že hypotéza H je pravdivá.

Bez ohľadu na to, aká veľká je pravdepodobnosť, že piata kolóna nerobí sabotáže, čiže P(~E|H1), nemôže byť taká veľká ako pravdepodobnosť, že *neexistujúca piata kolóna* nerobí sabotáže, čiže P(~E|H2). Pozorovanie neprítomnosti sabotáže teda zvyšuje pravdepodobnosť, že piata kolóna neexistuje.

Neprítomnosť sabotáže *nedokazuje*, že piata kolóna neexistuje. Neprítomnosť *dôkazu* nie je *dôkazom* neprítomnosti. V logike A -> B, „ak A, tak B", nie je to isté ako ~A -> ~B, „ak nie A, tak nie B".

Ale v teórii pravdepodobnosti, neprítomnosť *indície* je vždy *indíciou* neprítomnosti. Ak je E binárna udalosť a P(H|E) > P(H), čiže vidieť E zvyšuje pravdepodobnosť H, potom P(H|~E) < P(H), čiže nevidieť E znižuje pravdepodobnosť H. P(H) je váženým priemerom P(H|E) a P(H|~E), takže nevyhnutne leží medzi nimi. Ak vám čokoľvek z tohto znie nejasne, prečítajte si Intuitívne vysvetlenie bayesovského uvažovania.

V drvivej väčšine okolností z bežného života, príčina nemusí spoľahlivo vykazovať znaky svojej prítomnosti, ale neprítomnosť príčiny bude tieto znaky vykazovať ešte zriedkavejšie. Neprítomnosť pozorovania môže byť silnou indíciou neprítomnosti alebo slabou indíciou neprítomnosti, podľa toho, s akou pravdepodobnosťou daná príčina spôsobuje pozorovanie. Neprítomnosť pozorovania, ktoré by aj tak bolo veľmi zriedkavé (aj keď ho alternatívna hypotéza nedovoľuje vôbec), je veľmi slabou indíciou neprítomnosti (ale aj tak je to indícia). Toto je klam „medzier v zázname skamenelín" – skameneliny sa tvoria iba zriedkavo; je zbytočné roztrubovať neprítomnosť veľmi zriedkavého pozorovania, keď bolo zaznamenaných tak veľa silných pozitívnych pozorovaní. Ale ak nie sú absolútne *žiadne* pozitívne pozorovania, je čas znepokojovať sa; preto Fermiho paradox.

Tvoja sila racionalistu je tvoja schopnosť byť viac zmätený výmyslom než skutočnosťou; ak dokážeš rovnako dobre vysvetliť hocijaký výsledok, máš nulové vedomosti. Sila modelu nie je v tom,

48 Robyn M. Dawes, *Rational Choice in An Uncertain World*, [Rozumná voľba v neistom svete] 1st ed., ed. Jerome Kagan (San Diego, CA: Harcourt Brace Jovanovich, 1988), 250-251.

čo *môže* vysvetliť, ale v tom, čo *nemôže*, lebo iba zákazy obmedzujú očakávanie. Ak si nevšimneš, kedy tvoj model robí indíciu nepravdepodobnou, je to akoby si nemal žiaden model, a tiež akoby si nemal indície, mozog, ani oči.

⁂→

28. Zákon zachovania očakávanej indície

Friedrich Spee von Langenfeld, kňaz, ktorý počúval spovede odsúdených bosoriek, napísal v roku 1631 *Cautio Criminalis* („obozretnosť v kriminálnych prípadoch“), kde uštipačne opísal rozhodovací strom na odsudzovanie obvinených z bosoráctva: Ak bosorka viedla zlý a nesprávny život, je vinná; ak viedla dobrý a správny život, aj to je dôkaz, pretože bosorky sa pretvarujú a pokúšajú sa vyzerať zvlášť cnostne. Keď je žena uväznená: ak sa bojí, dokazuje to jej vinu; ak sa nebojí, dokazuje to jej vinu, pretože bosorky zvyčajne predstierajú nevinu a tvária sa statočne. Alebo keď počuje, že bola udaná z bosoráctva, môže sa pokúšať ujsť alebo zostať; ak utekala, dokazuje to jej vinu; ak zostala, diabol ju zadržal, takže nemohla odísť.

Spee robil spovedníka mnohým bosorkám; mal teda príležitosť vidieť *každú* vetvu obviňovacieho stromu, kde bez ohľadu na to, *čo* obvinená z bosoráctva povedala alebo urobila, vždy to bolo použité ako dôkaz proti nej. V každom jednotlivom prípade ste však počuli iba jednu z dvoch možných vetiev. Toto je dôvod, prečo si vedci vopred zapisujú svoje experimentálne predpovede.

Nemôžete mať jedno i druhé – z hľadiska teórie pravdepodobnosti, nielen férovosti. Pravidlo, že „neprítomnosť indície je indíciou neprítomnosti“ je špeciálnym prípadom všeobecnejšieho zákona, ktorý by som pomenoval Zákon zachovania očakávanej indície: Očakávaná *priemerná* výsledná pravdepodobnosť po zohľadnení indície sa musí rovnať pôvodnej pravdepodobnosti.

$$P(H) = P(H,E) + P(H,\sim E)$$

$$P(H) = P(H|E) \times P(E) + P(H|\sim E) \times P(\sim E)$$

Preto každému očakávaniu indície zodpovedá očakávanie rovnakej protiindície v opačnom smere.

Ak očakávate silnú pravdepodobnosť, že uvidíte slabú indíciu jedným smerom, musí to byť vyvážené slabým očakávaním videnia silnej indície opačným smerom. Ak ste si veľmi istí svojou teóriou a preto očakávate, že uvidíte výsledok zodpovedajúci vašej hypotéze, môže to posilniť váš názor iba o máličko (už je blízky 1); avšak nečakané zlyhanie vašej predpovede by malo (musí) dať vašej sebadôvere silný úder. Musíte očakávať, že budete mať *v priemere rovnakú* istotu ako na začiatku. Inými slovami, samotné *očakávanie*, že uvidíte indíciu – predtým než ju naozaj uvidíte – by nemalo zmeniť vaše pôvodné názory. (Opäť, ak vám toto nie je intuitívne zrejmé, pozrite si Intuitívne vysvetlenie bayesovského rozmýšľania.)

Ak teda tvrdíte, že „žiadna sabotáž“ je indíciou *pre* existenciu piatej kolónie Američanov japonského pôvodu, musíte tomu zodpovedajúco tvrdiť, že vidieť sabotáž by bolo argumentom *proti* existencii piatej kolónie. Ak tvrdíte, že „dobrý a správny život“ je indíciou, že žena je bosorka, potom zlý a nesprávny život musí byť indíciou, že nie je bosorka. Ak tvrdíte, že Boh odmieta prejaviť svoju existenciu, aby testoval vieru ľudí, potom zázraky opísané v Biblii musia byť argumentmi proti existencii Boha.

To neznie celkom správne, však? Venujte pozornosť tomuto pocitu, že *je to trochu nasilu*, tomu tichému hlásku v pozadí vašej mysle. Je to dôležité.

Pre skutočného bayesiána je nemožné hľadať indície, ktoré *potvrdzujú* nejakú teóriu. Neexistuje žiaden plán, ktorý by ste mohli vymyslieť, žiadna chytrá stratégia, žiaden prefíkaný nástroj, pomocou ktorého by ste mohli legitímne očakávať, že vaša dôvera v daný výrok bude (*v priemere*) vyššia než predtým. Môžete hľadať indície iba na to, aby ste teóriu *testovali*, ale nie aby ste ju potvrdili.

→ http:///lesswrong.com/lw/ih/absence_of_evidence_is_evidence_of_absence/

Toto uvedomenie dokáže veľmi odbremeniť vašu myseľ. Nemusíte sa znepokojovať ohľadom toho, ako interpretovať všetky možné experimentálne výsledky, aby potvrdzovali vašu teóriu. Nemusíte sa zaťažovať plánovaním, ako z *každej* štipky indície vyťažiť potvrdenie pre vašu teóriu, pretože viete, že každému očakávaniu indície zodpovedá očakávanie rovnakej protiindície v opačnom smere. Ak sa pokúsite oslabiť protiindíciu možného „abnormálneho" pozorovania, môžete to urobiť iba oslabením podpory „normálneho" pozorovania, v úplne rovnakej miere v opačnom smere. Je to hra s nulovým súčtom. Bez ohľadu na to, ako špekulujete, ako argumentujete, ako strategizujete, nemôžete očakávať, že výsledný plán hry posunie (v priemere) vaše názory konkrétnym smerom.

Môžete sa teda posadiť a uvoľniť, zatiaľ čo čakáte na príchod indície.

...ľudská psychológia je *taká* domotaná.

*→

29. Spätný pohľad znehodnocuje vedu

Tento úryvok z Meyersovej knihy *Skúmanie sociálnej psychológie*[49] sa oplatí prečítať celý.

Cullen Murphy, redaktor *The Atlantic*, povedal, že spoločenské vedy neprinášajú „žiadne myšlienky či závery, ktoré by sa nedali nájsť v [hociktorej] encyklopédii citátov... Deň za dňom sociológovia idú do sveta. Deň za dňom zisťujú, že správanie ľudí je viacmenej také, ako by ste čakali."

Samozrejme, celé toto „čakanie" je spätný pohľad. (Skreslenie spätného pohľadu: Osoby, ktoré poznajú správnu odpoveď na otázku, pripisujú omnoho väčšiu pravdepodobnosť tomu, že „by boli" uhádli túto odpoveď, v porovnaní s osobami, ktoré musia hádať bez poznania správnej odpovede.)

Historik Arthur Schlesinger, Jr. zavrhol vedecké štúdie skúseností vojakov z 2. svetovej vojny ako „nemotornú ukážku" zdravého rozumu. Napríklad:

1. Vojaci s vyšším vzdelaním mali väčšie problémy prispôsobiť sa než menej vzdelaní vojaci. (Intelektuáli boli menej pripravení na bojový stres než ľudia z ulice.)
2. Vojaci z juhu USA zvládali horúce podnebie South Sea Island lepšie než vojaci zo severu USA. (Južania sú viac zvyknutí na horúce počasie.)
3. Bieli vojaci sa viac snažili byť povýšení na poddôstojníkov než čierni vojaci. (Roky útlaku majú dopad na motiváciu k úspechu.)
4. Černosi z juhu dávali prednosť bielym dôstojníkom z juhu pred bielymi dôstojníkmi zo severu (pretože dôstojníci z juhu mali väčšiu skúsenosť a zručnosť v interakcii s černochmi).
5. Kým trvali boje, vojaci viac túžili po návrate domov, než keď vojna skončila. (Počas boja vojaci vedeli, že sú v smrteľnom nebezpečenstve.)

Koľko z týchto zistení si myslíte, že by ste *boli vedeli* predpovedať? 3 z 5? 4 z 5? Sú tam prípady, kde by ste boli predpovedali opak – kde by váš model dostal úder? Na chvíľku sa zamyslite, než budete pokračovať...

...

V tejto ukážke (od Paula Lazarsfelda prostredníctvom Meyersa) boli všetky horeuvedené zistenia *opakom* toho, čo sa naozaj zistilo.[50] Koľkokrát ste si mysleli, že váš model dostal úder? Koľkokrát ste priznali, že by ste sa boli mýlili? Tak taký dobrý bol váš model skutočnosti. Miera vašej sily racionalistu je vaša schopnosť byť viac prekvapený fikciou než skutočnosťou.

Samozrejme, pokiaľ som výsledky neotočil ešte raz. Čo si myslíte?

→ http://lesswrong.com/lw/ii/conservation_of_expected_evidence/

49 David G. Meyers, *Exploring Social Psychology* [Skúmanie sociálnej psychológie] (New York: McGraw-Hill, 1994), 15–19.

50 Paul F. Lazarsfeld, „The American Solidier—An Expository Review," Public Opinion Quarterly 13, no. 3 (1949): 377–404.

Máte pocit, že vaše myšlienkové postupy v tejto chvíli, keď *naozaj neviete* správnu odpoveď, sú iné než myšlienkové procesy, ktoré ste použili na racionalizovanie ľubovoľnej strany „známej“ odpovede?

Daphna Baratz dala vysokoškolákom dvojice domnelých zistení, jedno pravdivé („Počas prosperity ľudia míňajú väčšie percento svojho príjmu než počas krízy“) a jedno, ktoré bolo opakom pravdy.[51] V oboch prípadoch študenti ohodnotili údajné zistenie ako to, čo „by boli predpovedali“. Dokonalý štandardný omyl spätného pohľadu.

Ktorý vedie ľudí k záveru, že nepotrebujeme vedu, pretože by to všetko „boli predpovedali“.

(Presne tak, ako by ste boli predpovedali, však?)

Spätný pohľad nás vedie k systematickému podceňovaniu prekvapivosti vedeckých zistení, najmä objavov, ktorým *rozumieme* – tých, ktoré nám pripadajú skutočné, ktoré si vieme spätne dosadiť do svojich modelov sveta. Ak sa vyznáte v neurológii alebo fyzike a čítate si novinky z tejto oblasti, pravdepodobne podceňujete aj prekvapivosť zistení v týchto oblastiach. Toto neférovo znehodnocuje príspevok výskumníkov; a čo je horšie, zabraňuje vám to všimnúť si, keď vidíte indíciu, ktorá nezapadá do toho, čo by ste *naozaj* boli očakávali.

Musíme vynaložiť vedomé úsilie na to, aby sme boli *dostatočne* šokovaní.

*→

51 Daphna Baratz, *How Justified Is the „Obvious“ Reaction?* (Stanford University, 1983).
→ http://lesswrong.com/lw/im/hindsight_devalues_science/

D: Tajomné odpovede

30. Falošné vysvetlenia

Kde bolo, tam bolo, bola raz jedna učiteľka fyziky. Jedného dňa zvolala svojich žiakov do triedy a ukázala im široký štvorcový kovový plát vedľa horúceho radiátora. Žiaci položili ruky na plát a zistili, že strana pri radiátore je studená a odvrátená strana je horúca. A učiteľka povedala: *Prečo si myslíte, že je to tak?* Niektorí žiaci tipovali na prúdenie vzduchu, iní tipovali, že plát obsahuje zvláštne kovy. Vymysleli veľa tvorivých vysvetlení; nikto sa neznížil k tomu, aby povedal: „Neviem“ alebo: „Zdá sa mi to nemožné.“

Správna odpoveď znela, že pred vstupom žiakov do triedy, učiteľka otočila plát naopak.[52]

Zamyslime sa nad žiakom, ktorý horúčkovito koktá: „No, možno je to kvôli tepelnej vodivosti a takým veciam?“ Pýtam sa: jej táto odpoveď pravý názor? Tieto slová sa ľahko vyznávajú – vyslovujú hlasným precíteným hlasom. Ale dokážu tieto slová ovplyvňovať očakávanie?

Uvažujme o tomto nevinnom slovíčku „kvôli“, ktoré sa nachádza pred „tepelnej vodivosti“. Uvažujme o *ďalších* veciach, ktoré by sme zaň mohli napísať. Mohli by sme povedať, napríklad, „kvôli flogistonu“ alebo „kvôli mágii“.

„Mágia!“ vykríknete. „To nie je *vedecké* vysvetlenie!“ Isteže, vety „kvôli tepelnej vodivosti“ a „kvôli mágii“ ľahko rozoznáme ako patriace do rôznych *literárnych žánrov*. „Tepelná vodivosť“ je niečo, čo môže povedať povedzme Spock v *Star Treku*, zatiaľ čo „mágia“ je niečo, čo povie Giles v *Buffy, Lovkyni Upírov*.

My, ako bayesiánci, si však nevšímame literárne žánre. Pre nás je podstatou modelu jeho vplyv na očakávanie. Ak poviete „tepelná vodivosť“, aké vnemy vás to vedie *očakávať*? Za bežných okolností vás to vedie k očakávaniu, že ak položíte ruku na stranu plátu pri radiátore, budete vám táto strana pripadať teplejšia než vzdialená stranu. Ak „kvôli tepelnej vodivosti“ dokáže zároveň vysvetliť, prečo vám strana pri radiátore pripadá *chladnejšia*, potom dokáže vysvetliť prakticky *hocičo*.

A v tejto chvíli už všetci vieme (aspoň dúfam, že vieme), že keď viete rovnako dobre vysvetliť ľubovoľný výsledok, máte nulovú vedomosť. „Kvôli tepelnému prúdeniu“, použité týmto spôsobom, je maskovaná hypotéza maximálnej entropie. Z hľadiska očakávania je izomorfná povedaniu „mágia“. Znie to ako vysvetlenie, ale nie je.

Predstavte si, že by sme namiesto hádania merali teplotu kovového plátu v rôznych bodoch v rôznom čase. Keby sme videli kovový plát vedľa radiátora, za normálnych okolností by sme očakávali, že bodové teploty budú zodpovedať rovnováhe rovnice rozptylu vzhľadom na hraničné podmienky dané prostredím. Možno by ste nevedeli presnú teplotu prvého meraného bodu, ale po odmeraní prvých bodov – nie som dosť fyzikálne zdatný, aby som vedel, koľkých presne – by ste mohli urobiť výborný odhad zvyšku.

Skutočný majster umenia používania čísel na obmedzenie očakávania hmotných javov – „fyzik“ – by urobil pár meraní a povedal: „Tento plát bol v rovnováhe s prostredím pred dva a pol minútami, bol otočený, a teraz sa opäť približuje k rovnováhe.“

Omyl týchto žiakov nebol len v tom, že nedokázali obmedziť očakávanie. Ich hlbší omyl spočíval v domnení, že robia fyziku. Povedali slovo „kvôli“, nasledované slovami, ako hovorieva Spock v *Star Treku*, a mysleli si, že tým vstúpili do magistéria vedy.

Veru nie. Iba presunuli svoju mágiu z jedného literárneho žánru do druhého.

*→

—

52 Hľadajte „heat conduction.“ Zdroj: Joachim Verhagen, http://web.archive.org/web/20060424082937/http://-www.nvon.nl/scheik/best/diversen/scijokes/scijokes.txt, archívna verzia, 27. október 2001.

→ http://lesswrong.com/lw/ip/fake_explanations/

31. Hádanie učiteľovho hesla

Za mladi som čítal populárne fyzikálne knihy ako od Richarda Feynmana *QED: Zvláštna teória svetla a hmoty*. Vedel som, že svetlo sú vlny, zvuk sú vlny, hmota sú vlny. Vo veku deväť rokov som bol hrdý na svoju vedeckú gramotnosť.

Keď som bol starší a začal som si čítať *Feynmanove lekcie z fyziky*, našiel som poklad s názvom „vlnová rovnica“. Dokázal som pochopiť odvodenie tej rovnice, ale ani dodatočne som nevedel uvidieť jej pravdivosť na prvý pohľad. Tak som nad tou vlnovou rovnicou rozmýšľal s prestávkami tri dni, dokiaľ som neuvidel, aká je trápne samozrejmá. A keď som ju konečne pochopil, uvedomil som si, že po celý čas, čo som prijímal poctivé ubezpečenia fyzikov, že svetlo sú vlny, zvuk sú vlny, hmota sú vlny, nemal som najmenšie tušenie, čo slovo „vlna“ znamená pre fyzika.

Máme inštinktívny sklon myslieť si, že keď fyzik povie: „svetlo sa skladá z vĺn“ a učiteľ povie: „Z čoho sa skladá svetlo?“ a žiak povie: „Z vĺn!“, že ten žiak povedal pravdivý výrok. To je predsa férové, nie? Keď uznáme odpoveď „vlny“ ako správnu od fyzika, nebolo by neférové odmietnuť ju od žiaka? Odpoveď „Vlny!“ je určite buď *pravdivá* alebo *nepravdivá*, však?

A to je opäť jeden zvyk zo školy, ktorý sa treba odučiť. Slová nemajú definíciu sami od seba. Keď počujem slabiky „bo-bor“ a predstavím si veľkého hlodavca, je to fakt o stave mojej mysle, nie fakt o slabikách „bo-bor“. Postupnosť slabík „skladá sa z vĺn“ (alebo „kvôli tepelnej vodivosti“) nie je *hypotéza*, je to len vzor vibrácií putujúci vzduchom, alebo machuľa na papieri. Môže sa v mysli niekoho *spojiť* s hypotézou, ale samo od seba to nie je správne ani nesprávne. V škole vám však učiteľ dá hviezdičku za *vyslovenie* „skladá sa z vĺn“, čo musí byť správna odpoveď, pretože učiteľ počul, ako z fyzik vyludzoval tie isté zvukové vibrácie. Keďže slovné správanie (vyslovené alebo napísané) je to, čo vám prinesie hviezdičku, žiaci si začínajú myslieť, že slovné správanie má pravdivostnú hodnotu. Napokon, svetlo sa buď skladá z vĺn alebo nie, však?

A to vedie k ešte horšiemu zlozvyku. Predstavte si, že vám učiteľka predloží mätúci problém týkajúci sa kovového plátu vedľa radiátora; vzdialená strana je na dotyk teplejšia než strana pri radiátore. Učiteľka sa opýta: „Prečo?“ Ak poviete: „Neviem,“ *nemáte* šancu získať hviezdičku – ani len čiarku za aktivitu na hodine. Ale, počas tohto polroka táto učiteľka použila slovné spojenia „kvôli tepelnému prúdeniu“, „kvôli tepelnej vodivosti“ a „kvôli vyžarovaniu tepla“. Jedno z nich je asi to, čo učiteľka chce. Poviete: „No, možno kvôli tepelnej vodivosti?“

Toto nie je *hypotéza o* kovovom pláte. To nie je ani pravý názor. Je to pokus *uhádnuť učiteľkine heslo*.

Aj keď si predstavíte symboly rovnice rozptylu (matematiku riadiacu vedenie tepla), neznamená to, že ste vytvorili *hypotézu o* kovovom pláte. Toto nie je škola; my netestujeme vašu pamäť, či dokážete zapísať rovnicu rozptylu. Toto je bayesovské remeslo; my hodnotíme vaše očakávania vnemov. Ak *použijete* rovnicu rozptylu, odmeriate niekoľko bodov teplomerom a potom skúsite predpovedať, čo teplomer povie pri nasledujúcom meraní, potom je to jednoznačne spojené s vnemami. Dokonca aj keď si žiak iba predstavuje, ako niečo *prúdi* a preto prikladá merač k studenšej strane plátu v snahe odmerať, kam teplo išlo, potom sa tento myšlienkový obraz prúdenia spája so skúsenosťou; riadi očakávanie.

Ak *nepoužívate* rovnicu rozptylu – vkladaním čísel a získavaním výsledkov, ktoré riadia vaše očakávanie konkrétnych vnemov – potom toto spojenie medzi mapou a územím akoby bolo preťaté nožom. Čo zostalo, nie je názor, ale slovné správanie.

V školskom systéme je to všetko o slovnom správaní, či už napísanom na papieri alebo vyslovenom nahlas. Za slovné správanie dostanete hviezdičku alebo prepadnete. Súčasťou odvykania od tohto zlozvyku je uvedomovanie si rozdielov medzi vysvetlením a heslom.

Vyzerá toto príliš drsne? Keď čelíte mätúcemu kovovému plátu, nemôže „tepelná vodivosť?“ byť prvým krokom k nájdeniu odpovede? Možno, ale iba ak nespadnete do pasce myslenia si, že hľadáte heslo. Čo ak nemáte žiadneho učiteľa, ktorý by vám povedal, že je to zle? Potom si môžete myslieť, že „svetlo sú wakalixy“ je dobré vysvetlenie, že „wakalixy“ je správne heslo. To sa stalo mne, keď som

mal deväť rokov – nie preto, že by som bol hlúpy, ale pretože toto sa stáva *automaticky*. Takto ľudia myslia, pokiaľ nie sú vycvičení, aby do tejto pasce *nepadli*. Ľudstvo zostávalo v takýchto jamách zaseknuté celé tisícročia.

Možno, keby sme žiakov vycvičili, že *slová sa nepočítajú, iba ovládače očakávania*, žiaci by sa *nezasekávali* na: „Tepelná vodivosť? Nie? Žeby tepelné prúdenie? Ani to nie?“ Možno *potom* by myšlienka „tepelná vodivosť“ viedla k skutočne užitočnej ceste, ako:

- „Tepelná vodivosť?“
- Ale to je len fráza – čo to znamená?
- Rovnica rozptylu?
- Ale to sú len symboly – ako ich použijem?
- K akému očakávaniu ma vedie použitie rovnice rozptylu?
- Určite ma to nevedie k očakávaniu, že strana kovového plátu, ktorá je odvrátená od radiátora, bude na dotyk teplejšia.
- Všímam si, že som zmätený. Možno sa tá blízka strana iba *zdá* studenšou, pretože je vyrobená z izolujúcejšieho materiálu a prenáša menej tepla na moju ruku? Skúsim tú teplotu odmerať...
- Okej, to nebolo ono. Môžem skúsiť overiť, či rovnica rozptylu vôbec platí pre tento kovový plát? Či teplo *prúdi* tak, ako zvyčajne, alebo sa tu deje niečo iné?
- Mohol by som priložiť merač k plátu a skúsiť sledovať, ako sa teplo postupne šíri...

Ak *nie* sme dosť prísni ohľadom toho, že: „No, možno kvôli tepelnému prúdeniu?“ môže byť falošným vysvetlením, žiak sa veľmi pravdepodobne zasekne na nejakom hesle ako wakalixy. *Toto sa stáva automaticky, stávalo sa to ľudstvu ako celku po celé tisícročia.*

*

32. Veda ako uniforma

Upútavka na film *X-Men* obsahuje hlas, ktorý hovorí: „V každej ľudskej bytosti... sa nachádza genetický kód... pre mutáciu.“ Očividne môžete pomocou mutácie získať všemožné šikovné schopnosti. Mutantka Storm, napríklad, má schopnosť vrhať blesky.

Prosím ťa, drahý čitateľ, aby si zvážil biologické ústrojenstvo potrebné na generovanie elektriny; biologické adaptácie potrebné na zabránenie zraneniu vlastnou elektrinou; a kognitívne obvody potrebné na presne naladenú kontrolu bleskov. Keby sme v skutočnosti pozorovali nejaký organizmus, ktorý získal všetky tieto schopnosti *v jednej generácii* ako výsledok *mutácie*, naprosto by to vyvrátilo neodarwinovský model prirodzeného výberu. Bolo by to horšie než nájsť skamenelého králika v predkambriu. Keby evolučná teória *naozaj* dokázala vysvetliť Storm, bola by schopná vysvetliť čokoľvek a všetci vieme, čo by z toho vyplývalo.

Komix *X-Men* používa pojmy ako „evolúcia“, „mutácia“ a „genetický kód“ iba aby sa umiestnil do toho, čo považuje za vedecký *literárny žáner*. Čo ma na tom najviac desí, je predstava, koľko ľudí, najmä v médiách, chápe vedu *len* ako literárny žáner.

Stretávam sa s ľuďmi, ktorí veľmi jednoznačne veria v evolúciu a vysmievajú sa z hlúposti kreacionistov. A napriek tomu nemajú predstavu o tom, čo teória evolučnej biológie dovoľuje a zakazuje. Budú hovoriť o „ďalšom kroku v evolúcii ľudstva“, akoby sa sem prirodzený výber dostal podľa nejakého plánu. Alebo ešte horšie, budú hovoriť o niečom úplne mimo oblasti evolučnej biológie, napríklad o vylepšenom dizajne počítačových čipov, alebo o delení korporácií, alebo o ľuďoch, ktorí sa nahrajú do počítača, a budú *toto* volať „evolúcia“. Keby evolučná biológia mohla zahŕňať toto, mohla by zahŕňať všetko.

Pravdepodobne väčšina ľudí, ktorí *veria v* evolúciu, používa slovné spojenie „kvôli evolúcii“, pretože chcú byť súčasťou vedeckého davu – viera ako vedecká uniforma, ako keď nosíte laboratórny plášť. Keby vedecký dav namiesto toho používal frázu „kvôli inteligentnému dizajnu“, rovnako veselo by používali aj to – pre ich ovládače očakávania by v tom nebol žiaden rozdiel. Povedať „kvôli evolúcii“

namiesto „kvôli inteligentnému dizajnu“ *pre nich* nezakazuje Storm. Jediným účelom pre nich je identifikovať sa so skupinou.

Stretávam sa s ľuďmi, ktorí sú celkom ochotní baviť sa o myšlienke umelej inteligencie hlúpejšej než človek alebo dokonca aj umelej inteligencie mierne múdrejšej než človek. Zoznámte ich s pojmom silne nadľudskej umelej inteligencie a oni sa náhle rozhodnú, že je to „pseudoveda→“. Nie je to preto, že by si mysleli, že majú teóriu inteligencie, ktorá im dovoľuje vypočítať teoretickú hornú hranicu sily optimalizačného procesu. Namiesto toho si spájajú silnú nadľudskú UI s *literárnym žánrom* apokalyptickej literatúry; zatiaľ čo UI bežiacu v malej korporácii si spájajú s literárnym žánrom časopisu *Wired*. Nehovoria na základe modelu myslenia. Neuvedomujú si, že nejaký model *potrebujú*. Neuvedomujú si, že veda *je o* modeloch. Ich zdrvujúca kritika spočíva výlučne *v porovnaní s apokalyptickou literatúrou*, namiesto povedzme známych zákonov, ktoré zakazujú takýto výsledok. Chápu vedu *iba* ako literárny žáner alebo ako skupinu, do ktorej patria. Táto uniforma im nepripadá ako laboratórny plášť; nie je to ten futbalový tím, ktorému fandia.

Existuje nejaká časť vedy, na ktorú ste *hrdí*, že v ňu veríte a napriek tomu tento názor nevyužívate pracovne? Mali by ste sa sami seba opýtať, ktoré budúce vnemy táto viera *zakazuje*, aby sa vám stali. To je suma toho, čo ste vstrebali a stalo sa to naozaj vašou súčasťou. Všetko ostatné sú pravdepodobne heslá a uniformy.

*→

33. Falošná kauzalita

Flogiston bol odpoveďou 18. storočia na elementárny oheň gréckych alchymistov. Zapáľte drevo a nechajte ho horieť. Čo je to ten oranžovo svetlý „oheň“? Prečo sa drevo premieňa na popol? Chemici 18. storočia na obe otázky odpovedali: „flogiston“.

...a to je celé; skrátka, ich odpoveď bola: „Flogiston.“

Flogiston uniká z horiacej hmoty ako viditeľný oheň. Ako flogiston uniká, horiaca hmota stráca flogiston a stáva sa z nej popol, „skutočná hmota“. Oheň v uzavretej nádobe zhasne, pretože vzduch sa nasýti flogistonom a už ho viac neprijme. Uhlie zanecháva po spálení málo zvyškov, pretože je to takmer čistý flogiston.

Teória flogistonu sa samozrejme nedala použiť na *predpovedanie* výsledku chemickej reakcie. Najprv ste sa pozreli na výsledok, potom ste použili teóriu flogistonu, aby ste ho *vysvetlili*. Nebolo to tak, že by teoretickí flogistonici predpovedali, že oheň v uzavretej nádobe zhasne; namiesto toho zapálili oheň v nádobe, videli ako zhasol a potom povedali: „Vzduch musel byť nasýtený flogistonom.“ Nemohli ste použiť teóriu flogistonu na povedanie, čo by ste *nemali* uvidieť; dokázala vysvetliť všetko.

To bolo v raných časoch vedy. Dlho si nikto neuvedomil, že je to problém. Falošné vysvetlenia *neznejú* falošne. To je to, čo ich robí nebezpečnými.

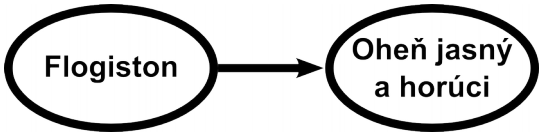
Moderný výskum naznačuje, že ľudia rozmýšľajú o príčinách a následkoch pomocou niečoho ako orientované acyklické grafy (DAGy) bayesovských sietí. Pretože pršalo, chodník je mokrý; pretože je chodník mokrý, šmýka sa.

Z tohto dokážeme odvodiť – alebo v bayesovskej sieti presne spočítať pravdepodobnosť – že ak sa chodník šmýka, pravdepodobne pršalo; ale ak už vieme, že chodník je mokrý, nová informácia, že chodník sa šmýka, nám nepovie nič nové o tom, či pršalo.

Prečo je oheň horúci a jasný keď horí?

→ http://lesswrong.com/lw/io/is_molecular_nanotechnology_scientific/

→ http://lesswrong.com/lw/ir/science_as_attire/



Znie to ako vysvetlenie. Je to *zapísané* v rovnakom formáte kognitívnych údajov. Ľudská myseľ však nedokáže automaticky rozoznať, kedy má príčina neobmedzujúcu šípku na následok. Čo je horšie, vďaka skresleniu spätného pohľadu sa nám zdá, že príčina obmedzovala následok, aj keď bola iba upravená, aby s následkom ladila.

Čo je zaujímavé, naše moderné chápanie pravdepodobnostného uvažovania o kauzalite dokáže presne opísať, čo robili teoretickí flogistonici nesprávne. Jednou z hlavných inšpirácií pre bayesovské siete bolo uvedomenie si problému dvojitého započítania indícií ak odvodzovanie rezonuje medzi následkom a príčinou. Povedzme napríklad, že som dostal kúsok nespoľahlivej informácie, že chodník je mokrý. To by malo spôsobiť, že si pomyslím, že je o čosi pravdepodobnejšie, že pršalo. Ale ak je pravdepodobnejšie, že pršalo, nie je potom pravdepodobnejšie, že chodník je mokrý? A neznamenalo by *to*, že je pravdepodobnejšie, že chodník sa šmýka? Ale ak sa chodník šmýka, asi je mokrý; a potom by som mal opäť zvýšiť pravdepodobnosť, že pršalo...

Judea Pearl použil ako metaforu algoritmus, ako sa počítajú vojaci v rade.[53] Predstavte si, že stojíte v rade a vidíte dvoch susedných vojakov, jedného pred vami a jedného za vami. To ste dokopy traja vojaci. Opýtate sa vojaka vedľa vás: „Koľkých vojakov vidíš *ty*?“ Obzrie sa a povie: „Troch.“ Tak to máme dokopy šesť vojakov. Očividne toto *nie* je správny spôsob, ako to robiť.

Rozumnejšie je opýtať sa vojaka pred vami: „Koľko vojakov je pred tebou?“ a vojaka za vami: „Koľko vojakov je za tebou?“ Otázka: „Koľko vojakov je pred tebou?“ sa dá ďalej odovzdať bezo zmätku. Ak som na začiatku radu, poviem výsledok: „1 vojak vpredu“. Osoba stojaca hneď za mnou dostane správu: „1 vojak vpredu“ a odovzdá výsledok: „2 vojaci vpredu“ vojakovi za ním. Podobne, každý vojak, ktorý dostane správu: „N vojakov vzadu“ od vojaka tesne za ním, ju odovzdá ako: „N+1 vojakov vzadu“ vojakovi pred ním. Koľko je vojakov celkovo? Spočítajte dve čísla, ktoré dostanete, pridajte jednotku za seba: toľko je v rade celkovo vojakov.

Pointa je, že každý vojak musí *oddelene* sledovať dve správy, správu o vojakoch vpredu a správu o vojakoch vzadu, a sčítať ich až na záver. Nikdy nepridávate vojakov z prijatej správy o vojakoch vzadu do správy o vojakoch vpredu, ktorú odovzdáte dozadu. A vôbec, celkové množstvo vojakov sa nikdy neodovzdáva ako správa – nikto ho nikdy nepovie nahlas.

Analogický princíp funguje pri dôkladnom pravdepodobnostnom uvažovaní o kauzalite. Ak sa niečo dozviete o tom, či pršalo, z nejakého *iného* zdroja než pozorovania, že chodník je mokrý, toto pošle správu dopredu z [Prší] do [Chodník je mokrý] a zvýši naše očakávanie, že chodník je mokrý. Ak pozorujete, že chodník je mokrý, toto pošle správu dozadu nášmu názoru, že pršalo, a táto správa pokračuje z [Prší] do všetkých susedných uzlov *okrem* uzla [Chodník je mokrý]. Každý kus indície započítame presne raz; žiadna správa sa „neodráža“ dopredu a dozadu. Presne tento algoritmus nájdete v klasickej knihe Judeu Pearla: „Pravdepodobnostné uvažovanie v inteligentných systémoch: Siete s dôveryhodným odvodzovaním.¬“

Čo bolo teda zlé na teórii flogistonu? Keď pozorujeme, že oheň je horúci, uzol [Oheň] môže poslať správu dozadu do uzla [Flogiston], čím aktualizujeme svoj názor o flogistone. Ale ak sa stalo toto, nemôžeme to považovať za úspešnú predpoveď teórie flogistonu smerom dopredu. Správa by mala ísť iba jedným smerom a neodrážať sa naspäť.

Žiaľ, ľudia nepoužívajú presné algoritmy na aktualizáciu sietí názorov. Učíme sa o rodičovských uzloch z pozorovania potomkovských uzlov a predpovedáme potomkovské uzly podľa názoroch o rodičovských. Pritom však prísne neoddeľujeme záznamy o správach dozadu a správach dopredu. Pamätáme si len, že flogiston je horúci, čo *spôsobuje*, že oheň je horúci. Takže sa zdá, že teória flogistonu predpovedá horúcosť ohňa. Ešte horšie, vyzerá to, že *flogiston spôsobuje, že oheň je horúci*.

53 Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Mateo, CA: Morgan Kaufmann, 1988).

→ https://books.google.sk/books?id=k9VsqN24pNYC

Dokiaľ si nevšimnete, že sa tu nerobia žiadne predpovede *dopredu*, nebudete tento neobmedzujúci kauzálny uzol považovať za „falošný“. Je zakreslený úplne rovnakým spôsobom ako hocijaký iný uzol vo vašej sieti názorov. Vyzerá ako fakt, ako všetky ostatné fakty, ktoré poznáte: *Flogiston spôsobuje, že oheň je horúci.*

Správne navrhnutá UI by si tento problém všimla hneď. Nevyžadovalo by to ani špeciálny kód na tento účel, iba správne záznamy o sieti názorov. (Žiaľ, my ľudia nedokážeme prepísať svoj vlastný kód tak, ako by to dokázala správne navrhnutá UI.)

Hovoriť o „skreslení spätného pohľadu“ je iba netechnický spôsob, ako povedať, že ľudia presne nerozdeľujú správy dopredu a správy dozadu, čím umožnia, aby správy dopredu boli kontaminované správami dozadu.

Tí, čo kedysi dávno šli cestou flogistonu, sa nesnažili byť hlúpi. Žiaden vedec sa úmyselne nepokúša zaseknúť v slepej uličke. Sú aj vo *vašej* hlave nejaké falošné vysvetlenia? Ak sú, ručím vám za to, že nie sú označené ako „falošné vysvetlenia“, takže hľadanie kľúčového slova „falošný“ vo vašich myšlienkach ich neodhalí.

Vďaka skresleniu spätného pohľadu takisto nestačí skontrolovať, ako dobre vaša teória „predpovedá“ fakty, ktoré už poznáte. Musíte predpovedať zajtrajšok, nie včerajšok. Je to jediný spôsob, ako zamotaná ľudská myseľ môže mať istotu, že posiela čistú správu dopredu.

* →

34. Sémantické stopky

A dieťa sa opýtalo:

> Odkiaľ pochádza tento kameň?
>
> Odštiepil som ho z veľkého balvanu v strede osady.
>
> Odkiaľ pochádza ten balvan?
>
> Asi sa skotúľal z vysokej hory, ktorá sa týči nad našou dedinou.
>
> Odkiaľ pochádza tá hora?
>
> Rovnako ako všetky skaly: sú to kosti Ymira, prvotného obra.
>
> Odkiaľ pochádza prvotný obor Ymir?
>
> Z veľkej priepasti Ginnungagap.
>
> Odkiaľ pochádza veľká priepasť Ginnungagap?
>
> Túto otázku sa nikdy nepýtaj.

Vezmime si zdanlivý paradox Prvej Príčiny. Veda vytrasovala udalosti späť k Veľkému Tresku, ale prečo sa stal samotný Veľký Tresk? Dá sa pekne odpovedať, že samotný čas začal v okamihu Veľkého Tresku – že pred Veľkým Treskom nie je žiaden tok minút a hodín. Ale aj táto odpoveď predpokladá naše fyzikálne zákony, ktoré samotné majú zložitú štruktúru; to si žiada vysvetlenie. Odkiaľ pochádzajú fyzikálne zákony? Mohli by ste povedať, že sme všetci v počítačovej simulácii, ale potom táto počítačová simulácia musí bežať na fyzikálnych zákonoch nejakého iného sveta – a odkiaľ pochádzajú *tie* fyzikálne zákony?

Vtedy niektorí ľudia povedia: „Boh!“

Ako by si mohol niekto, hoci aj veľmi nábožný človek, pomyslieť, že toto *pomohlo* odpovedať na paradox Prvej Príčiny? Prečo sa automaticky neopýtať: „Odkiaľ pochádza Boh?“ Povedať: „Boh je bez príčiny“ alebo „Boh stvoril sám seba“ nás necháva v celkom rovnakej pozícii ako: „Čas začal

→ http://lesswrong.com/lw/is/fake_causality/

pri Veľkom Tresku.“ Akurát sa opýtame, prečo vôbec existuje celý ten metasystém alebo prečo niektoré udalosti môžu byť bez príčiny a iné nie.

Mojím cieľom tu nie je diskutovať o zdanlivom paradoxe Prvej Príčiny, ale pýtať sa, ako si niekto môže myslieť, že „Boh!“ *rieši* tento paradox. Povedať „Boh!“ je spôsob, ako patriť do svojho kmeňa, čo dáva ľuďom motív hovoriť to tak často, ako sa len dá – niektorí ľudia to dokonca odpovedia aj na otázku: „Prečo tento hurikán zasiahol New Orleans?“ Napriek tomu by som dúfal, že si ľudia všimnú, že pri *konkrétnej* hádanke o Prvej Príčine povedať „Boh!“ nepomáha. Tento paradox vďaka tomu nevyzerá o nič menej paradoxne, *aj keby to bola pravda*. Ako je možné, že si to niekto *nevšimne*?

Jonathan Wallace→ naznačil, že „Boh!“ slúži ako *sémantická stopka*→ – že to nie je výrokové tvrdenie, ale skôr kognitívna dopravná značka: nerozmýšľajte za tento bod. Povedať „Boh!“ nerieši tento paradox, ale vysiela kognitívny dopravný signál na zastavenie zrejmého pokračovania reťaze otázok a odpovedí.

Samozrejme, *vy* ako dobrý a poriadny ateista by ste to nikdy neurobili, však? Lenže „Boh!“ nie je *jediná* sémantická stopka, je to len zrejmý prvý príklad.

Transhumánne technológie – molekulárna nanotechnológia, pokročilá biotechnológia, umelá inteligencia, a tak ďalej – nastoľujú ťažké politické otázky. Akú rolu, ak vôbec nejakú, by mala hrať vláda pri dozore nad rodičovským výberom génov pre svoje dieťa? Majú mať rodičia možnosť úmyselne vybrať gén pre schizofréniu? Ak je zvýšenie inteligencie dieťaťa drahé, mala by vláda pomôcť zabezpečiť jeho dostupnosť, aby zabránila vzniku kognitívnych elít? Môžete navrhovať rôzne inštitúcie, ktoré budú odpovedať na tieto politické otázky – napríklad, že súkromné nadácie by mali poskytovať finančnú pomoc na zvyšovanie inteligencie – ale samozrejmá ďalšia otázka je: „Bude táto inštitúcia efektívna?“ Ak sa budeme spoliehať, že právna zodpovednosť za výrobok zabráni firmám vytvoriť škodlivú nanotechnológiu, bude to naozaj *fungovať*?

Poznám človeka, ktorého odpoveď na každú z týchto otázok je: „Liberálna demokracia!“ A to je celé. To je jeho odpoveď. Ak mu položíte samozrejmú otázku ako: „Ako veľmi sa liberálnym demokraciám historicky darilo pri rovnako zložitých problémoch?“ alebo „Čo ak liberálna demokracia urobí niečo hlúpe?“, potom ste autokrat, libertopián, alebo iný druh veľmi veľmi zlého človeka. Nikto nesmie spochybňovať demokraciu.

Raz som tento spôsob uvažovania nazval: „božské právo demokracie“. Ale presnejšie by bolo povedať, že „Demokracia!“ preňho funguje ako sémantická stopka. Keby mu niekto povedal: „Prenechajme to firme Coca Cola!“, sám by položil samozrejmé ďalšie otázky: „Prečo? Čo s tým urobí firma Coca Cola? Prečo by sme jej mali dôverovať? Darilo sa jej v minulosti pri rovnako zložitých problémoch?“

Alebo predpokladajme, že by niekto povedal: „Američania mexického pôvodu intrigujú ako odstrániť všetok kyslík zo zemskej atmosféry.“ Pravdepodobne by ste sa opýtali: „Prečo by *toto* robili? Nemusia vari aj Američania mexického pôvodu dýchať? Fungujú vôbec Američania mexického pôvodu ako jednotná konšpirácia?“ Ak si nepoložíte tieto samozrejmé nasledujúce otázky keď niekto povie: „Korporácie intrigujú ako odstrániť kyslík“, potom je „Korporácie!“ vaša sémantická stopka.

Dajte si tu pozor, aby ste nevytvorili nový všeobecný protiargument na veci, ktoré sa vám nepáčia: „Ach, to je iba stopka!“ Žiadne slovo nie je samo osebe stopkou; otázka je, či má dané slovo takýto účinok na konkrétnu osobu. Mať voči niečomu silné emócie z toho ešte nerobí stopku. Ja nemám príliš rád teroristov, ani sa nebojím súkromného vlastníctva, to ale neznamená, že „Teroristi!“ alebo „Kapitalizmus!“ sú pre mňa kognitívne dopravné značky. (Slovo „inteligencia“ na mňa kedysi malo takýto účinok, ale už nie.) Poznávacím znamením sémantickej stopky je *neschopnosť zamyslieť sa nad samozrejmou ďalšou otázkou.*

*→

→ http://www.spectacle.org/yearzero/godvgod.html
→ http://www.spectacle.org/1095/stop1.html
→ http://lesswrong.com/lw/it/semantic_stopsigns/

35. Tajomné odpovede na tajomné otázky

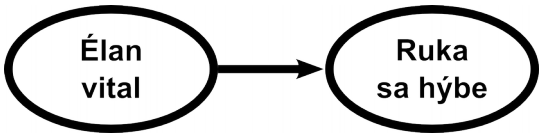
Predstavte si, že sa pozeráte na svoju ruku a neviete nič o bunkách, nič o biochémii, nič o DNA. Naučili ste sa čosi z anatómie pri pitve, takže viete, že vaša ruka obsahuje svaly; ale neviete, prečo sa svaly hýbu namiesto toho, aby iba ležali ako hlina. Vaša ruka je skrátka... vec... a z nejakého dôvodu sa hýbe podľa vašich rozhodnutí. Nie je to kúzlo?

> Živočíšne telo nefunguje ako termodynamický stroj … vedomie učí každého jednotlivca, že je do istej miery podriadený rozhodnutiam svojej vôle. Zdá sa teda, že pohyblivé tvory majú schopnosť okamžite vplývať na isté pohyblivé častice hmoty vo svojich telách silami, ktoré usmerňujú tieto častice, aby vykonali výsledný mechanický účinok... Pôsobenie živočíšneho alebo rastlinného života na hmotu je nekonečne ďaleko za hranicou akéhokoľvek doposiaľ podniknutého vedeckého bádania. Jeho schopnosť usmerňovania pohybov pohyblivých častíc, v každodenne pozorovanom zázraku našej ľudskej slobodnej vôle, i v raste generácie za generáciou rastlín z jediného semena, sa nekonečne odlišuje od ľubovoľného možného výsledku náhodného súbehu atómov... Moderní biológovia opäť začali čosi akceptovať a je to princíp života.
>
> --Lord Kelvin[54]

Toto bola teória *vitalizmu*; že tajomný rozdiel medzi živou a neživou hmotou vysvetľuje *élan vital* alebo *vis vitalis*. *Élan vital* napĺňal živú hmotu a spôsoboval, že sa hýbe tak, ako určuje vedomie. *Élan vital* sa zúčastňoval na chemických reakciách, ktoré púhe neživé častice nemohli podstúpiť – neskoršia Wöhlerova syntéza urey, zložky moču, bola veľkým úderom vitalistickej teórii, pretože ukázala, že púha *chémia* dokáže zopakovať produkt biológie.

Nazývať „élan vital“ vysvetlením, hoci aj falošným vysvetlením ako flogiston, pravdepodobne znamená brať ho príliš vážne. Fungoval v prvom rade ako zastavovač zvedavosti. Opýtali ste sa: „Prečo?“ a odpoveď znela: „Lebo élan vital!“

Keď ste povedali „Élan vital!“, *znelo* to, akoby ste vedeli, prečo sa vaša ruka hýbe. Mali ste v hlave malý kauzálny diagram, ktorý hovoril:



V skutočnosti ste však nevedeli o nič viac než predtým. Nevedeli ste napríklad povedať, či vaša ruka bude vytvárať teplo alebo pohlcovať teplo, pokiaľ ste si tento fakt už predtým nepozreli; ak nie, nevedeli ste ho predpovedať. Vaša zvedavosť vyzerala nasýtená, nebola však nakŕmená. Keďže môžete povedať: „Prečo? Lebo élan vital!“ na ľubovoľné možné pozorovanie, bolo to rovnaké dobré na vysvetlenie všetkých výsledkov, skrytá hypotéza maximálnej entropie, a tak ďalej.

Väčšie ponaučenie sa však skrýva za úctou vitalistov voči *élan vital*, ich ochota vyhlasovať ho za tajomstvo mimo všetku vedu. Pri stretnutí s drakom Neznámom, vitalisti nevytasili meče do útoku, ale podriadene sklonili hlavy. Boli hrdí na svoju nevedomosť, urobili z biológie *posvätné* tajomstvo a preto boli neochotní vzdať sa svojej nevedomosti, keď indície začali klopať na dvere.

Tajomstvo Života bolo *nekonečne ďaleko za hranicou vedy!* Nie iba *trochu* ďaleko, podotýkam, ale *nekonečne* ďaleko! Lord Kelvin bol iste mimoriadne emocionálne vzrušený z toho, že *niečo nevedel.*

Avšak nevedomosť existuje iba na mape, nie na území. Ak nerozumiem nejakému javu, je to fakt o mojom stave mysle, nie fakt o samotnom jave. Nejaký jav môže *pripadať* tajomný nejakej konkrétnej osobe. Neexistujú javy, ktoré sú tajomné sami osebe. Uctievať jav, pretože vyzerá tak zázračne tajomný, znamená uctievať svoju vlastnú nevedomosť.

Vitalizmus mal s flogistonom spoločnú chybu *zabalenia tajomstva ako podstaty*. Oheň bol tajomný a teória flogistonu zabalila toto tajomstvo do tajomnej podstaty nazvanej „flogiston“. Život bol posvätným tajomstvom, a vitalizmus zabalil toto tajomstvo do tajomnej podstaty nazvanej „élan vital“.

54 Silvanus Phillips Thompson, *The Life of Lord Kelvin* (American Mathematical Society, 2005).

Ani jedno z toho nepomohlo modelu sústrediť hustotu pravdepodobnosti – urobiť niektoré výsledky ľahšie vysvetliteľnými než iné. Toto „vysvetlenie“ iba zabalilo otázku do malej tvrdej nepriehľadnej čiernej guličky.

V Molierovej komédii lekár vysvetľuje schopnosť uspávadla tým, že obsahuje „dormitívnu potenciu“. Ten istý princíp. Je chybou ľudskej psychológie, že keď sa stretneme s tajomným javom, skôr predpokladáme tajomné základné podstaty než zložitý proces za tým.

Ale ešte hlbšou chybou je predpokladať, že *odpoveď* môže byť tajomná. Ak sa nejaký jav zdá tajomný, je to fakt o našom stave poznania, nie fakt o samotnom jave. Vitalisti videli tajomnú medzeru vo svojich vedomostiach, tak predpokladali tajomnú vec, ktorá túto medzeru zapĺňa. Keď tak robili, pomýlili si mapu s územím. Všetok zmätok existuje v mysli, nie v zabalených podstatách.

Toto je konečné a celkom všeobecné vysvetlenie, prečo sú v histórii ľudstva ľudia znovu a znovu šokovaní objavom, že neuveriteľne tajomná otázka má netajomnú odpoveď. Tajomnosť je vlastnosť otázok, nie odpovedí.

Preto nazývam teórie ako vitalizmus *tajomnými odpoveďami na tajomné otázky*.

Toto sú príznaky tajomných odpovedí na tajomné otázky:

- Po prvé, vysvetlenie je zastavovačom zvedavosti, nie ovládačom očakávania.
- Po druhé, hypotéza sa neskladá z pohyblivých častí – model nie je konkrétny zložitý mechanizmus, ale prázdna nehybná podstata či sila. Tajomná podstata alebo tajomná sila môže byť údajne tu alebo tam, môže spôsobiť toto alebo tamto; ale dôvod, prečo sa tajomná sila takto správa, je zahalený v nepreniknuteľnej prázdnote.
- Po tretie, tí, ktorí ponúkajú toto vysvetlenie, opatrujú svoju nevedomosť; hovoria hrdo o tom, ako tento jav vzdoruje bežnej vede alebo nie je ako púhe všedné javy.
- Po štvrté, *aj potom, čo dostanete odpoveď, daný jav je stále tajomstvom* a má takú istú zázračnú nevysvetliteľnosť, akú mal na začiatku.

*→

36. Márnosť emergencie

Zlyhania flogistonu a vitalizmu sú historickým spätným pohľadom. Odvážim sa ísť s kožou na trh a pomenovať nejakú *súčasnú* teóriu, ktorú považujem za analogicky pomýlenú?

Menujem *emergenciu* alebo *emergentné javy* – zvyčajne definované ako štúdium systémov, ktorých správanie na vyššej úrovni vzniká alebo „sa vynára“ z interakcie mnohých prvkov na nižšej úrovni. (Wikipédia: „spontánny vznik makroskopických vlastností a štruktúr zložitých systémov, ktoré nie je ľahké odvodiť z vlastností ich zložiek“.) Doslovne vzaté, tento popis sedí na každý jav v našom vesmíre nad úrovňou jednotlivých kvarkov, a to je časť problému. Predstavte si, že ukážete na krach burzy a poviete: „Toto nie je kvark!“ Znie vám to ako vysvetlenie? Nie? Potom by nemalo ani: „Je to emergentný jav!“

Je to podstatné meno „emergencia“, voči čomu protestujem, nie voči slovesu „vynárať sa“ [po anglicky: „emerge“]. Nie je nič zlé na povedaní „X sa vynára z Y“, kde Y je nejaký konkrétny podrobný model s vnútornými pohyblivými časťami. „Vyplýva z“ je ďalšie legitímne slovné spojenie, ktoré znamená presne to isté: Gravitácia vyplýva zo zakrivenia časopriestoru podľa konkrétneho matematického modelu všeobecnej relativity. Chémia vyplýva z interakcie medzi atómami podľa konkrétneho modelu kvantovej elektrodynamiky.

Teraz si predstavte, že by som povedal, že gravitácia je vysvetlená „vyplývavosťou“ alebo že chémia je „vyplývavostný jav“ a tvrdil, že toto je moje vysvetlenie.

Slovné spojenie „vynára sa z“ je prijateľné rovnako ako „vyplýva z“ alebo „je spôsobené“ sú prijateľné, ak za týmto slovným spojením nasleduje nejaký konkrétny model, ktorý hodnotíme podľa jeho vlastných kvalít.

→ http://lesswrong.com/lw/iu/mysterious_answers_to_mysterious_questions/

To však *nie* je spôsob, ako sa „emergencia“ bežne používa. „Emergencia“ sa bežne používa ako plnohodnotné vysvetlenie samo osebe.

Už si ani nepamätám, koľkokrát som som počul ľudí povedať napríklad: „Inteligencia je emergentný jav!“ akoby toto vysvetľovalo inteligenciu. Toto použitie spĺňa všetky body zoznamu pre tajomnú odpoveď na tajomnú otázku. Čo ste sa dozvedeli, keď si poviete, že inteligencia je „emergentná“? Nemôžete robiť žiadne nové predpovede. Neviete nič o správaní myslí v skutočnom svete, čo ste nevedeli predtým. Znie to, akoby ste verili v nový fakt, ale neočakávate žiadne odlišné výsledky. Vaša zvedavosť vyzerá nasýtená, nie však nakŕmená. Táto hypotéza nemá žiadne pohyblivé časti – nemá podrobný vnútorný model, ktorým sa dá pohybovať. Tí, ktorí ponúkajú hypotézu „emergencie“, vyznávajú svoju nevedomosť ohľadom vnútorných častí a sú na ňu hrdí; stavajú vedu „emergencie“ do protikladu k ostatným vedám všedných vecí.

A aj potom, čo dostanete odpoveď: „Prečo? Lebo emergencia!“, *daný jav stále zostáva tajomný* a má tú istú posvätnú nepreniknuteľnosť, ako mal na začiatku.

Zábavné cvičenie je vynechať prídavné meno „emergentný“ z každej vety, v ktorej sa nachádza a pozrieť, či tá veta hovorí niečo iné.

- *Predtým:* Ľudská inteligencia je emergentným výsledkom signalizácie neurónov.
- *Potom:* Ľudská inteligencia je výsledkom signalizácie neurónov.
- *Predtým:* Správanie kolónie mravcov je emergentným výsledkom interakcií mnohých jednotlivých mravcov.
- *Potom:* Správanie kolónie mravcov je výsledkom interakcií mnohých jednotlivých mravcov.
- *Ešte lepšie:* Kolónia sa skladá z mravcov. Dokážeme úspešne predpovedať niektoré vlastnosti správania kolónie použitím modelov, ktoré zahŕňajú iba jednotlivé mravce, bez globálnych premenných pre kolóniu, čo ukazuje, že rozumieme, ako toto správanie kolónie vzniká zo správania mravcov.

Ďalšie zábavné cvičenie je nahradiť slovo „emergentný“ slovom, ktoré ľudia pôvodne používali ako vysvetlenie pred vynálezom emergencie:

- *Predtým:* Život je emergentný jav.
- *Potom:* Život je zázračný jav.
- *Predtým:* Ľudská inteligencia je emergentným výsledkom signalizácie neurónov.
- *Potom:* Ľudská inteligencia je zázračným výsledkom signalizácie neurónov.

Nesprostredkuje vari každé tvrdenie celkom rovnaké množstvo vedomostí o správaní daného javu? Nezodpovedá vari každá hypotéza celkom rovnakej množine výsledkov?

Slovo „emergencia“ sa stalo veľmi obľúbené, rovnako ako slovo „zázrak“ bolo kedysi obľúbené. Slovo „emergencia“ má pre ľudskú psychológiu hlbokú príťažlivosť, z toho istého dôvodu. Slovo „emergencia“ je úžasne ľahkým vysvetlením a znie dobre, keď sa povie; dáva vám posvätné tajomstvo, ktoré môžete uctievať. Emergencia je obľúbená, *pretože* je to odpadové jedlo pre zvedavosť. Pomocou emergencie dokážete vysvetliť čokoľvek a presne to aj ľudia robia; lebo je to úžasný pocit vedieť vysvetliť veci. Ľudia sú stále ľuďmi, aj keď absolvujú pár hodín vedy na vysokej. Keď raz nájdu spôsob, ako sa striasť okov ustálenej vedy, pustia sa do rovnakých machinácií ako ich predkovia, oblečú ich do literárneho žánra „vedy“, ale je to stále psychológia toho istého živočíšneho druhu.

*→

37. Nehovorte „zložitosť“

Kde bolo, tam bolo...

Toto je príbeh môjho prvého stretnutia s Marcellom, s ktorým som neskôr rok pracoval na teórii UI; v tej dobe som ho však ešte neprijal ako učňa. Vedel som, že súťažil na celoštátnej úrovni matematických

→ http://lesswrong.com/lw/iv/the_futility_of_emergence/

a počítačových olympiád, čo stačilo na upútanie mojej pozornosti; nevedel som však, či sa dokáže naučiť rozmýšľať o UI.

Požiadal som Marcella, aby povedal, ako si myslí, že by UI mohla objaviť spôsob riešenia Rubikovej kocky. Nie vopred naprogramovaným spôsobom, čo je triviálne, ale skôr ako by UI mohla sama objaviť zákony Rubikovho sveta a zistiť, ako ich využívať. Ako by UI mohla *sama vynájsť* pojmy ako „operátor“ alebo „makro“, ktoré sú kľúčom na vyriešenie Rubikovej kocky?

V nejakom bode diskusie Marcello povedal: „No, myslím si, že UI potrebuje mať zložitosť, aby vedela urobiť X a zložitosť, aby vedela urobiť Y...“

A ja som povedal: „Nehovor ‚*zložitosť*‘.“

Marcello: „Prečo nie?“

Povedal som: „Zložitosť by nikdy nemala byť cieľom sama osebe. Môžeš potrebovať použiť nejaký konkrétny algoritmus, ktorý pridáva určité množstvo zložitosti, ale samotná zložitosť pre zložitosť veci iba komplikuje.“ (Myslel som pritom na všetkých ľudí, ktorých som počul obhajovať názor, že internet sa „prebudí“ a stane sa UI, keď sa stane „dostatočne zložitým“.)

A Marcello povedal: „Ale musí existovať *nejaké* množstvo zložitosti, ktoré to dokáže.“

Na chvíľu som zavrel oči a skúšal som rozmýšľať, ako to celé vysvetliť slovami. Mne osobne povedať „zložitosť“ jednoducho *pripadalo* ako nesprávny krok v tanci UI. Nikto nedokáže rozmýšľať dosť rýchlo na to, aby slovne zvážil každú vetu vo svojom prúde vedomia; to by si vyžadovalo nekonečnú rekurziu. Uvažujeme v slovách, ale náš prúd vedomia je usmerňovaný pod úrovňou slov, nacvičenými zvyškami minulých vhľadov a drsných skúseností...

Povedal som: „Čítal si Technické vysvetlenie technického vysvetlenia?“

„Áno,“ povedal Marcello.

„Dobre,“ povedal som. „Povedať ‚zložitosť‘ nekoncentruje tvoju masu pravdepodobnosti.“

„Aha,“ povedal Marcello, „ako ‚emergencia‘. Ha. Takže... by som teraz mal rozmýšľať o tom, ako by sa X mohlo naozaj stať...“

A vtedy som si pomyslel: „***Tento** je možno vzdelávateľný.*“

Zložitosť nie je zbytočný pojem. Sú k nemu matematické definície, ako Kolmogorovská zložitosť alebo Vapnikovská-Červonenkisovská zložitosť. Dokonca aj na intuitívnej úrovni sa často oplatí myslieť na zložitosť – potrebujete zvážiť zložitosť hypotézy a rozhodnúť sa, či je „príliš zložitá“ vzhľadom na podporujúce indície alebo pozrieť na dizajn a skúsiť ho zjednodušiť.

Ale pojmy nie sú užitočné alebo neužitočné sami osebe. Iba ich *použitie* je správne alebo nesprávne. V tanečnom kroku, ktorý sa Marcello pokúšal urobiť, sa snažil dostať vysvetlenie zadarmo, dostať niečo za nič. Je to mimoriadne častý chybný krok, prinajmenšom v mojej oblasti. Môžete sa zapojiť do diskusie o umelej všeobecnej inteligencii a sledovať ľudí, ktorí robia to isté, naľavo i napravo, znova a znova – nepretržite preskakujú veci, ktorým nerozumejú, neuvedomujúc si, čo práve robia.

Stane sa to v okamihu: umiestnite neovládajúci kauzálny uzol za čosi tajomné; kauzálny uzol, ktorý vyzerá ako vysvetlenie, ale nie je. Táto chyba sa vyskytuje pod úrovňou slov. Nevyžaduje si žiadnu mimoriadnu charakterovú vadu; je spôsob, ako ľudia štandardne myslia už od dávnych čias.

To, čomu sa musíte vyhnúť, je *preskočenie tej tajomnej časti*; musíte zastať pri tajomstve a čeliť mu priamo. Existuje veľa slov, ktorými sa tajomstvá dajú preskakovať a niektoré z nich by boli legitímne v inom kontexte – napríklad „zložitosť“. Ale podstatou chyby je toto *preskočenie*, bez ohľadu na to, aký kauzálny uzol je za ním. Toto preskočenie nie je myšlienka, iba mikromyšlienka. Musíte venovať dôkladnú pozornosť, aby ste sa pri nej prichytili. A keď sa nacvičíte vyhýbať preskakovaniu, stane sa to vecou inštinktu, nie slovného uvažovania. Musíte *cítiť*, ktoré časti vašej mapy sú stále prázdne, a čo je dôležitejšie, venovať tomuto pocitu pozornosť.

Predpokladám, že v akadémii je obrovský tlak na zametanie problémov pod koberec, aby ste mohli prezentovať článok s dojmom úplnosti. Viac slávy zožnete za napohľad úplný model, ktorý zahŕňa pár „emergentných javov“, než za explicitne neúplnú mapu s nápismi „fakt neviem, ako táto časť funguje“ alebo „a potom sa stane zázrak“. V časopise by vám možno ten druhý článok ani neprijali, pretože ktovie, či tie neznáme kroky nie sú vlastne tie, kde sa všetko to zaujímavé deje. A áno, občas sa stane, že všetky

tie nezázračné časti vašej mapy sa ukážu ako nedôležité. To je cena, ktorú občas platíte za vstup na neznáme územie a snahu riešiť problémy *postupne*. Ale v tom prípade je *ešte dôležitejšie vedieť*, kedy ešte nie ste na konci. Väčšinou sa ľudia na neznáme územie neodvažujú vstúpiť, pretože sa príliš boja, že premárnia svoj čas.

A ak zakladáte revolučnú novú firmu zameranú na UI, je tam ešte väčší tlak zamiesť problémy pod koberec; inak si musíte priznať, že zatiaľ neviete, ako zostrojiť UI a vaše súčasné životné plány sa rozsypú. Ale možno to teraz zbytočne zložito vysvetľujem, veď preskakovanie je u ľudí normálne; ak hľadáte príklady, sledujte ľudí, ktorí diskutujú o náboženstve alebo o filozofii alebo o duchovne alebo o nejakej vede, v ktorej neboli profesionálne vycvičení.

Marcello a ja sme si pri našej práci na UI vytvorili zvyk: keď sme narazili na niečo, čomu sme nerozumeli, a to bolo dosť často, povedali sme „zázrak“ - napríklad: „X zázrakom urobí Y“ - aby sme si sami pripomenuli, že *je tu nevyriešený problém, medzera v našom chápaní*. Je omnoho lepšie povedať „zázrak“ než „zložitosť“ alebo „emergencia“; tie druhé slová totiž vytvárajú ilúziu porozumenia. Múdrejšie je povedať „zázrak“ a nechať tak sám sebe značku, pripomienku práce, ktorú ešte bude treba dorobiť.

* →

38. Pozitívne skreslenie: Pozerajte do tmy

Učím v triede a píšem na tabuľu tri čísla: 2 – 4 – 6. „Myslím si pravidlo,“ hovorím, „ktoré sa týka trojíc čísel. Postupnosť 2 – 4 – 6, zhodou okolností, toto pravidlo spĺňa. Každý z vás má na lavici kôpku papierikov. Napíšte na papierik postupnosť troch čísel, a ja ju označím ‚Áno‘, ak spĺňa toto pravidlo, alebo ‚Nie‘, ak nespĺňa toto pravidlo. Potom môžete napísať ďalšiu trojicu čísel a opýtať sa, či tieto spĺňajú toto pravidlo a tak ďalej. Keď ste si istí, že poznáte to pravidlo, napíšte pravidlo na papierik. Môžete skúšať toľko trojíc, koľko len chcete.“

Tu je záznam otázok jedného študenta:

4, 6, 2	Nie
4, 6, 8	Áno
10, 12, 14	Áno

V tomto bode dotyčný študent napísal svoj odhad pravidla. Čo si *vy* myslíte, že je toto pravidlo? Chceli by ste vyskúšať ešte ďalšiu trojicu, a ak áno, ktorá by to bola? Pred ďalším čítaním sa na chvíľu zamyslite.

Uvedená úloha sa zakladá na klasickom experimente Petra Wasona, úloha 2 – 4 – 6. Hoci pokusné osoby v tejto úlohe zvyčajne vyjadrovali vysokú dôveru vo svoj odhad, iba 21 % študentov úspešne uhádlo experimentátorove skutočné pravidlo a neskoršie replikácie naďalej ukazujú mieru úspešnosti okolo 20 %.[55]

Táto štúdia sa volá „O neúspešnom eliminovaní hypotéz v pojmovej úlohe“. Pokusné osoby, ktoré skúšali vyriešiť úlohu 2 – 4 – 6, zvyčajne skúšajú vygenerovať *pozitívne* príklady namiesto *negatívnych* príkladov – používajú hypotetické pravidlo na vytvorenie reprezentatívneho príkladu a sledujú, či bude označený „Áno“.

Preto, keď si niekto vytvorí hypotézu „číslo sa zvyšuje o dve“, otestuje trojicu 8 – 10 – 12, dozvie sa, že vyhovuje a sebaisto oznámi toto pravidlo. Keď si niekto vytvorí hypotézu X – 2X – 3X, otestuje trojicu 3 – 6 – 9, dozvie sa, že vyhovuje a potom ohlási toto pravidlo.

V každom prípade je skutočné pravidlo to isté: tri čísla musia byť v rastúcom poradí.

→ http://lesswrong.com/lw/ix/say_not_complexity/

55 Peter Cathcart Wason, „On the Failure to Eliminate Hypotheses in a Conceptual Task, [O neúspešnom eliminovaní hypotéz v pojmovej úlohe]“ *Quarterly Journal of Experimental Psychology* 12, no. 3 (1960): 129–140, doi:10.1080/17470216008416717.

Ale aby ste objavili toto, museli by ste vytvárať trojice, ktoré by *nemali* vyhovovať, napríklad 20 – 23 – 26 a sledovať, či budú tieto označené „Nie“. Čo ľudia v tomto experimente nezvyknú robiť. V niektorých prípadoch si pokusné osoby vymyslia, „otestujú“ a oznámia pravidlá omnoho komplikovanejšie, než je skutočná odpoveď.

Tento kognitívny jav je zvyčajne zahrnutý pod „sklon potvrdzovať“. Mne sa však zdá, že jav testovania *pozitívnych* príkladov namiesto *negatívnych* by sme mali odlíšiť od javu zachovávania pôvodných názorov. Niekedy sa ako synonymum pre „sklon potvrdzovať“ používa „pozitívne skreslenie“, čo sa k tejto konkrétne chybe hodí omnoho lepšie.

Kedysi sa zdalo, že teória flogistonu dokáže vysvetliť, prečo plameň v uzavrenej nádobe zhasne (vzduch je presýtený flogistonom a už nedokáže prijať ďalší), ale teória flogistonu by rovnako dobre vedela vysvetliť, prečo plameň *nezhasne*. Aby ste si toto všimli, musíte hľadať negatívne príklady namiesto pozitívnych, pozerať na nuly namiesto jednotiek; čo ide proti tomu, čo experiment odhalil ako ľudský inštinkt.

Pretože podľa inštinktov ľudia žijú iba v polovici sveta.

Človeka možno učiť o pozitívnom skreslení celé dni a predsa ho v danej chvíli dokáže prehliadnuť. Pozitívne skreslenie nie je niečo, čo robíme kvôli logike alebo hoci len kvôli emocionálnej pripútanosti. Úloha 2 – 4 – 6 je „chladná“, logická, nie emocionálne „horúca“. A predsa je táto chyba sub-verbálna, na úrovni predstavivosti, inštinktívnej reakcie. Keďže problém nespočíva vo vytvorení vedomého pravidla, ktoré hovorí: „Mysli iba na pozitívne príklady“, nemožno ho vyriešiť slovným poznaním: „Mali by sme myslieť aj na pozitívne aj na negatívne príklady.“ Ktoré príklady vám automaticky naskakujú do hlavy? Musíte sa naučiť, bez slov, nie cik, ale cak. Musíte sa naučiť cuknúť smerom k nule, namiesto od nej.

Už hodnú dobu píšem o tom, že sila hypotézy je v tom, čo *nedokáže* vyriešiť, nie čo *dokáže* – že ak viete rovnako dobre vysvetliť hocijaký výsledok, máte nulové poznanie. Takže ak chcete zbadať vysvetlenie, ktoré nepomáha, nestačí pomyslieť na to, čo by dokázalo dobre vysvetliť – musíte hľadať aj výsledky, ktoré by *nedokázalo* vysvetliť, a toto je skutočná sila danej teórie.

Takže toto všetko som už povedal a potom som včera spochybnil užitočnosť „emergencie“ ako pojmu. Jeden diskutér uviedol supravodivosť a feromagnetizmus ako príklad emergencie. Odpovedal som, že nesupravodivosť a neferomagnetizmus sú tiež príkladmi emergencie a že práve v tom je háčik. Ale nebudem tohto diskutéra kritizovať! Napriek rozsiahlemu čítaniu o „sklone potvrdzovať“, sám som si nevšimol „chyták“ v úlohe 2 – 4 – 6, keď som o nej prvýkrát čítal. Je to subverbálna okamžitá reakcia, ktorú treba zmeniť výcvikom. Ja na sebe ešte stále pracujem.

Tak veľa z racionalistových schopností je pod úrovňou slov. Je to náročná práca pokúšať sa sprostredkovať toto Umenie pomocou článkov na blogu. Ľudia s vami budú súhlasiť, a potom, v nasledujúcej vete, nevedomky urobia niečo úplne opačné. Nesťažujem sa! Jedným z hlavných dôvodov, prečo sem píšem, je pozorovanie, čo moje slová *nedokázali* sprostredkovať.

Hľadáte práve teraz pozitívne príklady pozitívneho skreslenia, alebo si šetríte časť svojho hľadania na to, čo by ste podľa pozitívneho skreslenia vidieť *nemali*? Pozeráte sa smerom do svetla alebo do tmy?

*→

—

39. Zákonitá neistota

V knihe *Rozumná voľba v neistom svete* Robyn Dawes opisuje experiment, ktorý uskutočnil Tversky:[56,57]

→ http://lesswrong.com/lw/iw/positive_bias_look_into_the_dark/

56 Dawes, *Rational Choice in An Uncertain World*; Yaacov Schul and Ruth Mayo, „Searching for Certainty in an Uncertain World: The Difficulty of Giving Up the Experiential for the Rational Mode of Thinking,“ *Journal of Behavioral Decision Making* 16, no. 2 (2003): 93–106, doi:10.1002/bdm.434.

57 Amos Tversky and Ward Edwards, „Information versus Reward in Binary Choices,“ *Journal of Experimental Psychology* 71, no. 5 (1966): 680–683, doi:10.1037/h0023123.

> Koncom 1950-tych a začiatok 1960-tych rokov sa urobilo veľa psychologických pokusov, v ktorých žiadali pokusné osoby predpovedať výsledok udalosti, ktorá mala nejakú náhodnú zložku, ale jej základná miera sa predsa dala predpovedať -- pokusné osoby mali napríklad predpovedať, či nasledujúca otočená karta bude červená alebo modrá, v kontexte, kde 70 % kariet bolo modrých, ale postupnosť červených a modrých kariet bola celkom náhodná.
>
> V takejto situácii najväčší podiel úspechov prináša stratégia predpovedať tú častejšiu z udalostí. Napríklad, ak je 70 % kariet modrých, potom predpovedať modrú pri každej otázke dáva mieru úspechu 70 %.
>
> Pokusné osoby však namiesto toho mali sklon dávať zodpovedajúce pravdepodobnosti -- čiže predpovedať pravdepodobnejšiu udalosť s relatívnou frekvenciou, s akou sa vyskytovala. Napríklad mali sklon predpovedať v 70 % prípadov, že vyjde modrá karta, a v 30 % prípadov, že vyjde červená karta. Takáto stratégia dáva mieru úspechu 58 %, pretože pri výskyte modrej karty (čo sa stáva s pravdepodobnosťou 0,7) majú pokusné osoby pravdu v 70 % prípadov, a pri výskyte červenej karty (čo sa stáva s pravdepodobnosťou 0,3) majú pravdu v 30 % prípadov, a 0,7 × 0,7 + 0,3 × 0,3 = 0,58.
>
> V skutočnosti pokusné osoby predpovedali tú častejšiu udalosť s pravdepodobnosťou o trochu vyššou než s akou sa vyskytovala, ale nepriblížili sa k jej predpovedaniu v 100 % prípadov, dokonca ani keď ich platili za presnosť ich predpovedí... Napríklad, ak pokusným osobám platili päťcent za každú správnu predpoveď z tisíca pokusov... predpovedali [tú častejšiu udalosť] v 76 % prípadov.

Nemyslite si, že tento pokus sa týka nejakej drobnej chyby v stratégiách hazardných hier. Presne vykresľuje tú najdôležitejšiu myšlienku celej rozumnosti.

O tomto pokuse Dawes ďalej hovorí: „Napriek spätnej väzbe z tisíca pokusov, pokusné osoby nedokázali uveriť, že je to situácia, v ktorej *nedokážu* predpovedať.“

Lenže tá chyba musí byť ešte hlbšia než toto. Aj keby si pokusné osoby *mysleli*, že došli na nejakú hypotézu, nemusia podľa nej *naozaj tipovať*, aby ju otestovali. Môžu povedať: „Ak je *táto* hypotéza správna, nasledujúca karta bude červená“ - a potom jednoducho staviť na modrú.

Nezazlieval by som pokusnej osobe, že pokračuje vo vymýšľaní hypotéz - odkiaľ môže vedieť, že táto postupnosť je naozaj mimo jej schopností predpovedať? Ale zazlieval by som pokusnej osobe, že *na tieto tipy stávkuje*, hoci to nie je potrebné na získanie informácie, a doslova *stovky* predchádzajúcich tipov sa nepotvrdili.

Dokáže byť hoci aj človek *takto* prehnane sebaistý?

Mám podozrenie, že sa jedná o niečo jednoduchšie -- že pokusným osobám stratégia výlučnej modrej *jednoducho ani nenapadla*.

Ľudia vidia zmes prevažne modrých kariet a zopár červených, a predpokladajú, že optimálna stratégia stávkovania musí byť zmes prevažne modrých kariet a zopár červených.

Myšlienka, že pri neúplnej informácii sa *optimálna stratégia stávkovania nepodobá na typickú postupnosť kariet*, je *proti intuícii*.

Myšlienka, že optimálna stratégia je správať sa zákonite, napriek tomu, že prostredie obsahuje náhodné prvky, je *proti intuícii*.

Vyzerá to, akoby vaše správanie malo byť nepredpovedateľné, rovnako ako prostredie – ale nie! *Náhodný kľúč neodomkne náhodný zámok len vďaka tomu, že sú „oba náhodné“*.

Proti ohňu nebojujete ohňom, ale vodou. Lenže toto si vyžaduje krok navyše, nový pojem, ktorý sa priamo nespája so zadaním problému, a nie je to prvá myšlienka, ktorá vám napadne.

V dileme modrých a červených kariet nám naše čiastočné poznanie hovorí - v každom jednom kole - že najlepší tip je modrá. Táto rada nášho čiastočného poznania je rovnaká v každom jednom kole. Ak v 30 % prípadov ideme proti nášmu čiastočnému poznaniu a stavíme namiesto toho na červenú, bude sa

nám tak dariť horšie -- pretože teraz sme už vyslovene hlúpi a stávkujeme na to, čo vieme, že je menej pravdepodobný výsledok.

Keby ste *v každom* kole stavili na červenú, robili by ste to tak zle, ako sa len dá; boli by ste na 100 % hlúpi. Ak kvôli 30 % červených kariet stavíte na červenú v 30 % prípadov, potom sa robíte na 30 % hlúpi. Pri neúplnej informácii optimálna stratégia stávkovania nepripomína typickú postupnosť kariet.

Keď je vaše poznanie neúplné -- čo znamená, že sa vám zdá, že svet okolo vás má prvok náhody -- robiť náhodné činy nerieši daný problém. Náhodné konanie vás od cieľa vzďaľuje, nie približuje. Vo svete, ktorý už je plný hmly, zahodiť svoju inteligenciu veci iba zhorší.

Myšlienka, že optimálna stratégia môže byť *rozmýšľať zákonito, dokonca aj za podmienok neistoty*, je *proti intuícii*.

A preto racionalistov nie je veľa, pretože väčšina tých, ktorí vnímajú chaotický svet, sa pokúsi bojovať proti chaosu chaosom. Musíte urobiť krok navyše, a pomyslieť na niečo, čo vám priamo nevyskočí na myseľ, aby ste si predstavili bojovať proti ohňu niečím, čo samotné nie je oheň.

Počuli ste, ako neosvietení hovoria: „Rozumnosť funguje dobre pri jednaní s rozumnými ľuďmi, ale svet nie je rozumný.“ Lenže *ak zoči-voči nerozumnému súperovi odhodíte svoj vlastný rozum, nepomôže vám to*. Existujú zákonité formy myslenia, ktoré stále vytvárajú tú najlepšiu reakciu, dokonca aj zoči-voči súperovi, ktorí porušuje tieto zákony. Teória rozhodovania *nevzbĺkne* plameňom a nezomrie, keď sa stretne so súperom, ktorý neposlúcha teóriu rozhodovania.

Toto nie je o nič zrejmejšie než myšlienka staviť vždy na modrú, zoči-voči postupnosti modrých a červených kariet. Lenže každá stávka, ktorú urobíte na červenú, je v priemere strata, a podobne je aj každá odchýlka od Cesty vo vašom vlastnom myslení.

Koľko epizód *Star Treku* sme takto vyvrátili? Koľko teórií UI?

* →

40. Moja divoká a bezstarostná mladosť

Hovorí sa, že rodičia robia všetky tie veci, ktoré hovoria svojim deťom, aby ich nerobili, pretože práve preto vedia, že je lepšie ich nerobiť.

Kedysi dávno, v nepredstaviteľne vzdialenej minulosti, som bol oddaný Tradičný Racionalista, považoval som sa za zručného podľa ich merítok, nepoznal som však Bayesovu Cestu. Keď mladý Eliezer natrafil na tajomne vyzerajúcu otázku, prikázania Tradičnej Racionality mu nezabránili vymyslieť Tajomnú Odpoveď. Je to tá zďaleka najzahanbujúcejšia chyba v mojom živote a stále mnou mykne, keď na ňu pomyslím.

Čo bola moja tajomná odpoveď na tajomnú otázku? To nebudem popisovať, lebo by to bol dlhý a zložitý príbeh. Bol som mladý, púhy Tradičný Racionalista nepoznajúci učenie Tverskeho a Kahnemana. Poznal som Occamovu britvu, nie však klam konjunkcie[→]. Myslel som, že mi prejde, keď budem rozmýšľať zložito, v literárnom štýle zložitých myšlienok, ktoré som čítal vo vedeckých knihách, neuvedomujúc si, že správna zložitosť je možná len vtedy, keď je každý jej krok poriadne pripevnený. Dnes, jedna z hlavných rád, ktoré dávam ašpirujúcim mladým racionalistom, je: „Nepokúšaj sa o dlhé reťaze úvah alebo zložité plány.“

Netreba povedať nič viac než toto: Aj potom, čo som vymyslel svoju „odpoveď“, mi bol daný jav stále tajomný a mal tú istú vlastnosť úžasnej nepreniknuteľnosti, ako mal na začiatku.

Nemýľte sa, mladý Eliezer nebol hlúpy. Všetky chyby, ktorými sa mladý Eliezer previnil, dodnes robia vážení vedci vo vážených časopisoch. Aby sa pred nimi ochránil, potreboval by zručnosť presnejšiu, než akej sa kedy učil ako Tradičný Racionalista.

→ http://lesswrong.com/lw/vo/lawful_uncertainty/

→ http://en.wikipedia.org/wiki/Conjunction_fallacy

Vskutku, mladý Eliezer usilovne a svedomito nasledoval príkazy Tradičnej Racionality počas svojho zblúdenia z cesty.

Ako Tradičný Racionalista, mladý Eliezer dbal na to, aby jeho Tajomná Odpoveď odvážne predpovedala budúcu skúsenosť. Konkrétne, očakával som, že v budúcnosti neurológovia objavia, že neuróny využívajú kvantovú gravitáciu, ako hovoril Sir Roger Penrose. To si vyžadovalo, aby si neuróny udržiavali určitú úroveň kvantovej koherencie, a to bolo niečo, čo ste mohli hľadať a buď to nájsť alebo nenájsť. Buď niečo pozorujete alebo nie, však?

Moja hypotéza však nerobila žiadne *spätné* predpovede. Podľa Tradičnej Vedy, spätné predpovede sa nepočítajú – načo sa s nimi teda zaťažovať? Na druhej strane, pre Bayesovca, ak *dnes* hypotéza nemá lepšiu podmienenú pravdepodobnosť než „neviem“, vyvoláva to otázku, prečo *dnes* veríte niečomu komplikovanejšiemu než „neviem“. Nepoznal som však Bayesovu Cestu, takže som nemyslel na podmienené pravdepodobnosti ani zameriavanie hustoty pravdepodobnosti. Urobil som Falzifikovateľnú Predpoveď; vari nie toto je Zákon?

Ako Tradičný Racionalista, mladý Eliezer si dával pozor, aby neveril v mágiu, mysticizmus, uhlíkový šovinizmus, ani nič toho druhu. Hrdo som vyznával svoju Tajomnú Odpoveď: „Je to iba fyzika, rovnako ako zvyšok fyziky!“ Akoby ste z mágie mohli urobiť niečo iné než kognitívny izomorf mágie, ak ju pomenujete kvantová gravitácia. Nepoznal som však Bayesovu Cestu a nevidel som úroveň, na ktorej moja myšlienka bola izomorfná mágii. Prisahal som *vernosť* fyzike, ale to ma nezachránilo; čo už vie teória pravdepodobnosti o vernosti? Vyhýbal som sa všetkému, o čom mi Tradičná Racionalita povedala, že je zakázané, avšak čo zostalo, bola stále mágia.

Niet pochybnosti, že vernosť Tradičnej Racionalite mi pomohla dostať sa z jamy, ktorú som si sám vykopal. Keby som nebol Tradičným Racionalistom, mohol som byť *úplne* v kýbli. Ale Tradičná Racionalita stále nestačila na to, robiť veci *správne*. Iba ma viedla k iným chybám než boli tie, ktoré vyslovene zakazovala.

Keď rozmýšľam na tým, ako moje mladšie ja veľmi starostlivo nasledovalo pravidlá Tradičnej Racionality v procese získavania *nesprávnej* odpovede, vrhá to svetlo na otázku, prečo ľudia, ktorí si hovoria „racionalisti“, nevládnu svetu. Potrebujete *sakra veľké množstvo rozumnosti* než vás privedie k niečomu inému ako novým a zaujímavým chybám.

Tradičná Racionalita sa vyučuje skôr ako umenie než ako veda. Čítate si životopisy slávnych fyzikov opisujúcich lekcie, ktoré ich život naučil, a pokúšate sa robiť to, čo vám povedali, aby ste robili. Lenže vy ste nežili ich životy a polovica z toho, čo sa pokúšajú opísať, je inštinkt, ku ktorému boli vycvičení.

Tradičná Racionalita je zostrojená tak, že by bolo prijateľné, aby som strávil 30 rokov nad svojou hlúpou myšlienkou, pokiaľ ju jedného dňa úspešne falzifikujem, budem dosť poctivý k sebe samému ohľadom toho, čo moja teória predpovedala, prijmem jej vyvrátenie, keď príde, a tak ďalej. To stačí na to, aby západka Vedy cvakla vpred, ale je to trochu drsné voči ľuďom, ktorí premárnia 30 rokov svojho života. Tradičná Racionalita je chôdza, nie tanec. Je vytvorená, aby vás *jedného dňa* dostala k pravde a máte more času na privoniavanie ku kvetom pozdĺž cesty.

Tradiční Racionalisti sa môžu zhodnúť na tom, že sa nezhodnú. Tradiční Racionalisti nemajú ten *ideál*, že myslenie je presné umenie, v ktorom za daných indícií existuje iba jeden správny odhad pravdepodobnosti. V Tradičnej Racionalite máte povolené hádať a potom testovať vaše odhady. Skúsenosť ma však naučila, že ak niečo *neviete* a hádate, ukáže sa, že ste sa mýlili.

Bayesova Cesta je tiež nepresným umením, prinajmenšom v tej podobe, ako o nej rozprávam. Tieto články na blogu sú stále iba tápavými pokusmi vložiť do slov lekcie, ktoré by sa lepšie naučili skúsenosťou. Ale prinajmenšom je *založená* na matematike, plus experimentálnych indíciách z kognitívnej psychológie o tom, ako ľudia naozaj myslia. Možno to bude stačiť na dosiahnutie stratosfericky vysokej latky potrebnej na výcvik, ktorý vám umožní robiť to naozaj správne, namiesto púheho obmedzenia sa na nové zaujímavé chyby.

*→

41. Neučíme sa z histórie

Kde bolo, tam bolo, za mojej divokej a bezstarostnej mladosti, keď som nepoznal Bayesovu Cestu, dal som Tajomnú Odpoveď na tajomne vyzerajúcu otázku. Bola to postupnosť mnohých chýb, ale jedna chyba mi vyčnieva ako najkritickejšia: Moje mladšie ja si neuvedomilo, že *vyriešené tajomstvo by malo pôsobiť menej mätúco.* Pokúšal som sa vysvetliť Tajomný Jav – čo pre mňa znamenalo poskytnúť mu príčinu, zapasovať ho do prepojeného modelu skutočnosti. Prečo by toto malo urobiť daný jav menej Tajomným, ak je to jeho podstata? Pokúšal som sa *vysvetliť* Tajomný Jav, nie premeniť ho (nejakou nemožnou alchýmiou) na všedný jav, jav, ktorý by v prvom rade ani nevyzeral, že si vyžaduje nejaké nezvyčajné vysvetlenie.

Ako Tradičný Racionalista som poznal historické príbehy astrológov a astronómie, alchymistov a chémie, vitalistov a biológie. Ale tento Tajomný Jav bol iný. Bolo to niečo *nové*, niečo zvláštnejšie, niečo zložitejšie, niečo, čo bežná veda nedokázala vysvetliť celé stáročia...

...ako keby hviezdy a hmota a život neboli tajomstvami počas stoviek a tisícov rokov, od úsvitu ľudského myslenia až kým ich veda konečne nevyriešila...

Učíme sa o astronómii a chémii a biológii v školách a pripadá nám, že tieto veci *vždy* boli skutočným územím vedy, že *nikdy* neboli tajomné. Keď sa veda pokúsi zaútočiť na novú Veľkú Skladačku, deti danej generácie sú skeptické, pretože nikdy nevideli, že by veda vysvetlila niečo, čo im *pripadalo* tajomné. Veda je dobrá iba na vysvetľovanie *vedeckých* tém, ako sú hviezdy a hmota a život.

Myslel som si, že lekciou histórie je, že astrológovia a alchymisti a vitalisti mali vrodenú charakterovú vadu, sklon k mysticizmu, ktorý ich viedol k tajomným vysvetleniam netajomných vecí. Avšak, keby nejaký jav *naozaj bol* veľmi čudný, mohlo by čudné vysvetlenie byť namieste?

Až dodatočne, keď som začal vidieť všednú štruktúru vnútri tajomstva, som si uvedomil, v koho topánkach som stál. Až vtedy som si uvedomil, ako rozumne znel *svojho času* vitalizmus, aká *prekvapujúca* a *zahanbujúca* bola odpoveď vesmíru: „Život je všedný a nevyžaduje si čudné vysvetlenie."

Čítame si históriu, ale *nežijeme* v nej, *neprežívame* ju. Kiež by som bol ja *osobne* predpokladal astrologické tajomstvá a potom objavil newtonovskú mechaniku, predpokladal alchymistické tajomstvá a potom objavil chémiu, predpokladal vitalistické tajomstvá a potom objavil biológiu. Napadla by mi moja Tajomná Odpoveď a povedal by som si: *Na toto sa už opäť nachytať nedám.*

*→

42. Sprístupniť históriu

Existuje myšlienkový zvyk, ktorý by som nazval *logickým omylom zovšeobecňovania fiktívnej indície*, a ten si jedného dňa zaslúži samostatný článok na blogu. Novinári, ktorí napríklad hovoria o filme *Terminátor* v spravodajstve o UI, zvyčajne neberú *Terminátora* ako proroctvo alebo nemennú pravdu. Ale tento film sa spomína – je dostupný[¯] – akoby to bol ilustrujúci historický prípad. Akoby daný novinár videl, ako sa to odohralo na nejakej inej planéte, takže by sa to ľahko mohlo stať aj tu. Viac o tom je v kapitole 7 článku „Kognitívne skreslenia potenciálne ovplyvňujúce hodnotenie globálnych rizík".[58]

Existuje aj opačná chyba k zovšeobecňovaniu fiktívnej indície: nedostatočné ovplyvnenie *historickou* indíciou. Problém zovšeobecňovania fiktívnej indície je, že je to fikcia – nikdy sa to naozaj

→ http://lesswrong.com/lw/iy/my_wild_and_reckless_youth/

→ http://lesswrong.com/lw/iz/failing_to_learn_from_history/

→ http://en.wikipedia.org/wiki/Availability_heuristic

58 Eliezer Yudkowsky, „Cognitive Biases Potentially Affecting Judgment of Global Risks," in *Global Catastrophic Risks*, ed. Nick Bostrom and Milan M. Ćirković (New York: Oxford University Press, 2008), 91–119.

nestalo. Nepochádza z rovnakej distribúcie ako náš skutočný svet; fikcia sa systematickým spôsobom odlišuje od skutočnosti.[→] Avšak história sa *stala* a *mala* by byť dostupná.

V našom pravekom prostredí neexistovali kiná; čo ste videli na vlastné oči, to bola pravda. Je teda niečo čudné na tom, že fikcie, ktoré vidíme v živo sa pohybujúcich obrázkoch, majú na nás priveľký dopad? Naopak, veci, ktoré sa *naozaj stali*, stretávame iba ako atrament na papieri; stali sa, ale my sme ich nikdy *nevideli* stať sa. Nepamätáme sa, že by sa stali nám.

Táto opačná chyba je považovať históriu za púhy príbeh, spracovávať ju rovnakou časťou mozgu ako román, ktorý čítate. Môžete ústami povedať, že je to „pravda" a nie „fikcia", ale to neznamená, že vás to ovplyvní natoľko, ako by malo. Mnohé skreslenia obsahujú nedostatočné ovplyvnenie suchou, abstraktnou informáciou.

Kde bolo, tam bolo, dal som Tajomnú Odpoveď na tajomnú otázku, neuvedomujúc si, že robím presne tú istú chybu ako astrológovia hľadajúci tajomné vysvetlenia pre hviezdy, alebo alchymisti hľadajúci magické vlastnosti hmoty, alebo vitalisti predpokladajúci, že nejasný „élan vital" vysvetlí celú biológiu.

Keď som si nakoniec uvedomil, v koho topánkach som stál, bol to náhly šok z nečakaného spojenia s minulosťou. Uvedomil som si, že objav a skaza vitalizmu – o ktorých som čítal iba v knihách – sa *naozaj stali skutočným ľuďom*, ktorí ich prežívali viacmenej tak, ako som ja prežíval objav a skazu svojej vlastnej tajomnej odpovede. A tiež som si uvedomil, že keby som bol naozaj *zažil* minulosť – keby som žil počas minulých vedeckých revolúcií osobne, nie iba čítal o nich v historických knihách – pravdepodobne by som *nebol* zopakoval tie isté chyby. Neprišiel by som s *ďalšou* tajomnou otázkou; prvých tisíc lekcií by mi to ponaučenie vtĺklo do hlavy.

Preto (pomyslel som si), aby som dostatočne precítil silu histórie, mal by som skúsiť napodobniť myšlienky, ktoré by mal Eliezer *žijúci* počas histórie – mal by som skúsiť myslieť, akoby sa všetko, o čom som čítal v historických knihách, v skutočnosti stalo mne. (S primeraným zohľadnením skreslenia dostupnosti historických kníh – mal by som si pamätať, že som bol tisíckrát poddaným za každý raz, čo som bol kráľom.) Mal by som sa ponoriť do histórie, predstaviť si *život* v obdobiach, ktoré som videl iba ako atrament na papieri.

Prečo by som si mal pamätať prvý let bratov Wrightovcov? Nebol som tam. Ale ako racionalista, môžem sa opovážiť *nepamätať* si udalosť, ktorá sa naozaj stala? Je až taký veľký rozdiel medzi videním udalosti na vlastné oči – čo je v skutočnosti kauzálna reťaz zahŕňajúca odrazené fotóny, nie priame spojenie – a videním udalosti cez historickú knihu? Fotóny a historické knihy obe pochádzajú z kauzálnych reťazí od samotnej udalosti.

Musel som prekonať falošnú amnéziu narodenia sa v konkrétnej dobe. Musel som si pripomenúť – sprístupniť – *všetky* spomienky, nielen spomienky, ktoré zhodou okolností patrili mne a mojej vlastnej dobe.

Zem zrazu zostarla.

V mojej pôvodnej pamäti, Spojené Štáty vždy existovali – nikdy nebol čas, kedy by neboli žiadne Spojené Štáty. Nepamätal som sa, dovtedy, ako rástla Rímska ríša, ako priniesla mier a poriadok, ako vydržala toľko storočí, až som zabudol, že veci boli kedysi inak; a ako tá ríša predsa padla, barbari dobyli moje mesto, a poznanie, ktoré som zhromaždil, sa stratilo. Moderný svet sa pred mojimi očami stal krehkým; nebol to prvý moderný svet.

Tak veľa chýb, robených znovu a znovu a *znovu*, pretože som si nepamätal, že som ich urobil, v každej dobe, keď som nežil...

Pomyslite na to, že ľudia sa občas pýtajú, či je prekonávanie skreslení dôležité.

Nepamätáte sa, koľkokrát vás vaše skreslenia zabili? Nie? Všimol som si, že náhla strata pamäti často nasleduje po smrteľnej chybe. Ale verte mi, stalo sa to. Pamätám sa; nebol som tam.

Až nabudúce začnete pochybovať o čudnosti budúcnosti, spomeňte si, ako ste sa narodili v tlupe lovcov a zberačov pred desaťtisíc rokmi, keď nikto vôbec nepoznal Vedu. Spomeňte si, ako ste boli

---

→ http://www.overcomingbias.com/2007/07/tell-your-anti-.html

šokovaní, do jadra svojej bytosti, keď Veda vysvetlila veľké a strašné posvätné tajomstvá, ktoré ste kedysi tak veľmi uctievali. Spomeňte si, ako ste kedysi verili, že dokážete lietať, keď zjete tie správne huby, potom ste sklamane prijali, že lietať nikdy nebudete, a potom ste leteli. Spomeňte si, ako ste si vždy mysleli, že otroctvo je správne a potom ste zmenili názor. Nepredstavujte si, ako by ste túto zmenu *dokázali* predpokladať,[→] pretože to je strata pamäte. *Spomeňte si*, že v skutočnosti ste to neuhádli. Spomeňte si, ako sa storočie za storočím svet menil spôsobom, ktorý ste nečakali.

Možno vás potom bude menej šokovať, čo sa stane ďalej.

*→

43. Vysvetli / Uctievaj / Ignoruj?

Ako sa náš kmeň túla lúkami, hľadajúc ovocné stromy a korisť, z času na čas sa stane, že z oblohy padá voda.

„Prečo občas padá voda z oblohy?“ pýtam sa bradatého mudrca nášho kmeňa.

Na chvíľu sa zamyslí, lebo mu táto otázka nikdy predtým nenapadla, a potom povie: „Z času na čas duchovia oblohy bojujú a vtedy z oblohy kvapká ich krv.“

„Odkiaľ sú títo duchovia oblohy?“ pýtam sa.

Jeho hlas sa mení na šepot. „Sú starší než čas. Sú dávni, dávni.“

Keď prší a neviete prečo, máte niekoľko možností. Po prvé, môžete sa jednoducho nepýtať prečo – nesledovať túto otázku, alebo si ju v prvom rade ani nikdy nepomyslieť. To je možnosť Ignoruj, ktorú si bradatý mudrc vybral pôvodne. Po druhé, môžete sa pokúsiť vymyslieť nejaké vysvetlenie, to je možnosť Vysvetli, ako urobil bradatý muž v odpovedi na moju prvú otázku. Po tretie, môžete si vychutnávať pocit tajomnosti – to je možnosť Uctievaj.

Avšak, ako ste si iste všimli z tohto príbehu, vždy keď si vyberiete Vysvetli, v najlepšom prípade dostanete vysvetlenie ako sú „duchovia oblohy“. Lenže aj samotné toto vysvetlenie predmetom rovnakej voľby – Vysvetli, Uctievaj alebo Ignoruj? Vždy keď kliknete Vysvetli, veda chvíľu hrkoce, vráti vysvetlenie, a potom vyskočí ďalšie dialógové okno. Ako dobrí racionalisti cítime povinnosť stále klikať Vysvetli, ale vyzerá to ako cesta bez konca.

Kliknete Vysvetli na život a dostanete chémiu; kliknete Vysvetli na chémiu a dostanete atómy; kliknete Vysvetli na atómy a dostanete elektróny a jadrá; kliknete Vysvetli na jadrá a dostanete kvantovú chromodynamiku a kvarky; kliknete Vysvetli na vznik kvarkov a dostanete Veľký Tresk...

Môžeme kliknúť Vysvetli na Veľký Tresk a čakať, kým sa veda prehrkoce cez svoj proces a možno jedného dňa vráti úplne dobré vysvetlenie. Ale to opäť prinesie iba ďalšie dialógové okno. Ak teda budeme klikať dosť dlho, musíme sa dostať k *špeciálnemu* dialógovému oknu, k *novej* možnosti, k Vysvetleniu Ktoré Nepotrebuje Vysvetlenie, k miestu kde táto reťaz končí... a toto je možno jediné vysvetlenie, ktoré je hodno poznať.

Tak... a práve som klikol Uctievaj.

Nikdy nezabudnite, že existuje mnoho iných spôsobov uctievania okrem zapaľovania sviečok okolo oltára.

Keby som povedal: „Uf, to znie ako paradox. Ktovie, aké je riešenie tohto zdanlivého paradoxu?“ vtedy by som bol klikol Vysvetli, ktorému občas chvíľu trvá než vyprodukuje odpoveď.

A ak vám celá táto téma pripadá nedôležitá, nesúvisiaca, alebo by ste radšej rozmýšľanie o nej odložili na zajtra, potom ste klikli Ignoruj.

Vyberajte si svoju možnosť múdro.

*→

→ http://lesswrong.com/lw/il/hindsight_bias/

→ http://lesswrong.com/lw/j0/making_history_available/

→ http://lesswrong.com/lw/j2/explainworshipignore/

44. „Veda“ ako zastavovač zvedavosti

Predstavte si, že by som v plnom zábere televíznych kamier zdvihol ruky a zaspieval *abrakadabra* a spôsobil, že sa zrodí žiarivé svetlo v prázdnom vzduchu nad mojimi vystretými rukami. Predstavte si, že by som vykonal tento čin jasného nepomýliteľného čarodejníctva pod plným dozorom Jamesa Randiho a celej armády skeptikov. Myslím si, že väčšina ľudí by bola *pomerne zvedavá*, čo sa deje.

Teraz si však predstavte, že nejdem do televízie. Nechcem sa deliť o svoju moc, ani o pravdu za ňou. Chcem, aby moje kúzlo bolo tajné. Zároveň však chcem čarovať svoje kúzla kedykoľvek a kdekoľvek sa mi zachce. Chcem začarovať svoje žiarivé svetlo, aby som si mohol vo vlaku čítať knihu – a aby sa tomu nikto nečudoval. Existuje kúzlo na zastavenie zvedavosti?

Veru áno! Keď sa niekto opýta: „Ako si to urobil?“, jednoducho poviem: „Veda!“

Nie je to skutočné vysvetlenie, ale iba zastavovač zvedavosti. Nepovie vám, či to svetlo bude časom jasnejšie alebo bledšie, či zmení odtieň alebo sýtosť farby, a rozhodne vám to nepovie, ako si vyrobiť svoje vlastné podobné svetlo. V skutočnosti *neviete* o nič viac než ste vedeli predtým ako som povedal toto čarovné slovo. Avšak vy sa odvrátite, spokojný, že sa tu nedeje nič nezvyčajné.

Ešte lepšie, ten istý trik funguje aj s obyčajným vypínačom.

Prepnite vypínač a rozsvieti sa žiarovka. Prečo?

V škole vás naučia, že heslo k žiarovke je „Elektrina!“ Dúfam, že teraz sa už vyvarujete označenia žiarovky za „pochopenú“ na základe tohto. Umožní vám slovo „Elektrina!“ urobiť výpočet na očakávanie vnemov? Prinajmenšom sa treba naučiť veľa ďalších vecí. (Fyzici by mali ignorovať tento odsek a nahradiť si ho problémom z evolučnej teórie, kde je podstata teórie opäť vo výpočtoch, ktoré vie urobiť iba pár ľudí.)

Keby ste si mysleli, že žiarovka je *vedecky nevysvetliteľná*, upútalo by to *celú* vašu pozornosť. Odložili by ste všetko ostatné, čo ste práve robili, a sústredili by ste sa na žiarovku.

Ale čo znamená fráza „vedecky vysvetliteľné“? Znamená to, že niekto *iný* vie, ako žiarovka funguje. Keď vám povedia, že žiarovka je „vedecky vysvetliteľná“, neviete o nič viac, než ste vedeli predtým; neviete, či bude žiarovka časom jasnejšia alebo bledšia. Ale pretože niekto *iný* to vie, znižuje to vo vašich očiach hodnotu tohto poznania. Stanete sa menej zvedavý.

Niekto určite povie: „Keby bola tá žiarovka vedeckou záhadou, mohol by si jej skúmaním získať slávu a peniaze.“ Ale ja teraz nehovorím o chamtivosti. Nehovorím o kariérnej ambícii. Hovorím o čistej emócii zvedavosti – o pocite, že ma niečo zaujíma. Prečo by sa *vaša* zvedavosť mala zmenšiť vďaka tomu, že niekto *iný*, nie vy, vie, ako funguje žiarovka? Nie je to protivnosť? Že nestačí, aby ste to vedeli *vy*; ostatní ľudia musia byť v nevedomosti, inak z toho nebudete mať radosť?

Poznanie môže slúžiť aj iným veciam okrem zvedavosti, napríklad spoločensky užitočnej technológii. Na tieto inštrumentálne ciele je podstatné, či niekto iný v našom okolí už vie. Ale pre moju vlastnú zvedavosť, prečo by na tom malo záležať?

Navyše, predstavte si dôsledky toho, keby ste dovolili, aby „Niekto iný už vie odpoveď“ fungovalo ako zastavovač zvedavosti. Jedného dňa vojdete do svojej obývačky a uvidíte veľkého zeleného slona, ktorý sa vznáša vo vzduchu obklopený aurou strieborného svetla.

„Čo do čerta...?“ poviete.

A sponad slona zaznie hlas, ktorý povie: „NIEKTO INÝ UŽ VIE, PREČO JE TENTO SLON TU.“→

„Aha,“ poviete, „tak potom je mi to jedno,“ a idete do kuchyne.

Ja nepoznám všeobecnú zjednotenú teóriu vesmírnych zákonov fyziky. Neviem ani veľa o ľudskej anatómii, s výnimkou mozgu. Nevedel by na vlastnom tele ukázať, kde mám obličky, ani sa mi narýchlo nevybavuje, čo robí moja pečeň. (Nie som na to hrdý. Žiaľ, pri všetkej tej matematike, ktorú potrebujem naštudovať, pravdepodobne sa anatómiu tak skoro nenaučím.)

Mala by ma, pokiaľ ide o čistú *zvedavosť*, viac zaujímať moja nevedomosť ohľadom definitívnych zákonov fyziky než skutočnosť, že neviem veľa o tom, čo sa deje vnútri môjho vlastného tela?

→ http://godescalc.wordpress.com/2012/06/24/overlooked-elephant/

Keby som zdvihol ruky a začaroval svetelné kúzlo, zaujímalo by vás to. Mala by vás o niečo *menej* zaujímať samotná skutočnosť, že som zdvihol ruky? Keď vy zdvihnete svoju ruku a zamávate, tento čin koordinuje (okrem iného) váš mozoček. Stavím sa, že neviete, ako mozoček funguje. Ja viem trochu – hoci len hrubé detaily, ktoré nestačia na výpočet – ale čo z toho? Čo na tom záleží, ak *vy* neviete? Prečo by mal existovať dvojaký meter zvedavosti pri mágii a pohyboch ruky?

Pozrite sa na seba v zrkadle. Viete, na čo sa pozeráte? Viete, čo na vás pozerá spoza vašich očí? Viete, čo ste? Niektoré z týchto odpovedí Veda pozná a niektoré z nich Veda nepozná. Ale prečo by vašej zvedavosti malo záležať na tomto rozdiele, ak *vy* neviete?

Viete, ako fungujú vaše kolená? Viete, ako boli vyrobené vaše topánky? Viete, prečo váš monitor svieti? Viete, prečo je voda mokrá?

Svet okolo vás je plný otázok. Robte si priority, ak musíte. Ale nesťažujte sa, že krutá Veda pripravila svet o tajomstvá. S takýmto uvažovaním by som vedel spôsobiť, aby ste prehliadli slona vo vlastnej obývačke.

*→

45. Naozaj vo vás

Klasický článok Drewa McDermotta: „Umelá inteligencia stretáva prirodzenú hlúposť“ kritizuje programy UI, ktoré sa pokúšajú reprezentovať pojmy ako *šťastie je stav mysle* pomocou sémantickej siete:[59]

[ŠŤASTIE] –(JE)–> [STAV MYSLE]

A samozrejme *vo vnútri* uzla „ŠŤASTIE“ nie je nič; je to iba holý symbol LISPu so sugestívnym anglickým názvom.

Preto, hovorí McDermott: „Dobrým testom disciplinovaného programátora je pokúsiť sa namiesto takýchto kľúčov použiť vygenerované symboly a pozrieť, či ešte stále obdivuje svoj systém. Keby sme napríklad STAV MYSLE premenovali na G1073...“, mali by sme JE(ŠŤASTIE, G1073), „čo vyzerá o dosť pochybnejšie.“

Alebo, aby som trochu preformuloval túto myšlienku: Keby ste náhodnými symbolmi nahradili *všetky* sugestívne anglické názvy, nemali by ste najmenšiu šancu vydedukovať, čo vlastne to G1071(G1072, G1073) malo znamenať. Mal tým program UI reprezentovať hamburgery? Jablká? Šťastie? Ktovie. *Ak vymažete tie sugestívne anglické názvy, nedorastú naspäť.*

Predstavte si, že vám fyzik povie, že „Svetlo sú vlny“ a vy mu *veríte*. Teraz máte v hlave malú sieť, ktorá hovorí:

JE(SVETLO, VLNY)

Ak sa vás niekto opýta: „Čo je to svetlo?“, viete mu odpovedať: „Vlny!“

Ako hovorí McDermott: „Celý problém je, aby si počujúci všimol, čo mu povedali. Nie ‚pochopil‘, ale ‚všimol si‘.“ Predstavte si, že vám namiesto toho fyzik povie: „Svetlo sú malé krivolaké veci.“ (Mimochodom, to nie je pravda.) *Všimli* by ste si nejaký rozdiel v očakávaných vnemoch?

Ako si uvedomiť, že by ste nemali dôverovať svojmu domnelému poznaniu, že „svetlo sú vlny“? Jedna skúška, ktorú môžete použiť, je opýtať sa: „Dokázal by som túto vedomosť *obnoviť*, keby mi ju nejako vymazali z hlavy?“

Toto je podobná myšlienka, ako keď znečitateľníte mená sugestívne nazvaných symbolov LISPu vo vašom programe UI a skúsite, či niekto iný dokáže zistiť, na čo údajne „odkazovali“. Je to podobná myšlienka ako keď si všimnete, že Umelý Matematik, naprogramovaný, aby si zapamätal a vypisoval

Plus(Sedem, Šesť) = Trinásť

→ http://lesswrong.com/lw/j3/science_as_curiositystopper/

59 Drew McDermott, „Artificial Intelligence Meets Natural Stupidity,“ *SIGART Newsletter*, no. 57 (1976): 4–9, doi:10.1145/1045339.1045340.

nedokáže obnoviť túto vedomosť, ak mu ju vymažete z pamäte, dokiaľ ju nejaký človek opäť nezadá do databázy. Takisto, ak zabudnete, že „svetlo sú vlny“, nedokážete túto vedomosť získať naspäť žiadnym iným spôsobom než tým pôvodným – opýtať sa fyzika. Nedokážete túto vedomosť vytvoriť sami spôsobom, ktorým ju pôvodne vytvorili fyzici.

Rovnaká skúsenosť, ktorá nás vedie k formulovaniu názoru, spája tento názor s ostatným poznaním a zmyslovými vstupmi a pohybovými výstupmi. Ak vidíte bobra, ktorý obhrýza kmeň, potom viete, ako táto vec-ktorá-obhrýza-kmene vyzerá a dokážete ju rozoznať v budúcich situáciách, či už ju nazvete „bobor“ alebo nie. Ale ak získate svoje názory o bobroch od niekoho iného, kto vám povie fakty o „bobroch“, možno nebude vedieť rozoznať bobra, keď ho uvidíte.

To je to hrozné nebezpečenstvo skúšania *povedať* Umelej Inteligencii fakty, ktoré by sa nedokázala sama naučiť. To je aj to hrozné nebezpečenstvo skúšania *povedať* niekomu o fyzike, čo si nedokáže sám overiť. Lebo to, čo fyzici myslia slovom „vlna“ nie je „malá krivolaká vec“ ale čisto matematický pojem.

Ako hovorí Davidson, keby ste verili, že „bobry“ žijú v púšti, majú čisto bielu farbu a v dospelosti vážia 150 kíl, nemali by ste *o bobroch* žiadne názory, pravdivé ani nepravdivé. Váš názor o „bobroch“ nie je dosť dobrý na to, aby vôbec mohol byť nesprávny.[60] Ak nemáte dosť vnemov na obnovenie názorov, keď sa vymažú, máte dosť vnemov na to, aby ste tento názor vôbec s niečím spojili? Wittgenstein: „Koleso, ktorým možno točiť, aj keď sa nič iné nehýbe s ním, nie je súčasťou mechanizmu.“

Takmer hneď ako som začal čítať o UI – ešte skôr než som čítal McDermotta – som si uvedomil, že by bol *naozaj dobrý nápad* vždy si položiť otázku: „Ako by som obnovil túto vedomosť, keby mi ju nejako vymazali z hlavy?“

Čím hlbšie je vymazanie, tým prísnejší je test. Keby mi z hlavy vymazali všetky dôkazy Pytagorovej vety, vedel by som ju znovu dokázať? Myslím, že áno. Keby mi z hlavy vymazali všetky informácie o Pytagorovej vete, všimol by som si, že Pytagorovu vetu treba opäť dokázať? To už je ťažšie tvrdiť bez podkladov; ale keby ste mi dali pravouhlý trojuholník s odvesnami 3 a 4 a povedali mi, že dĺžka prepony sa dá vypočítať, myslím si, že by som to dokázal vypočítať, pokiaľ by som stále vedel zvyšok matematiky.

A čo samotný pojem *matematického dôkazu*? Keby mi o tom nikto nikdy nebol povedal, dokázal by som *toto* znovu objaviť na základe ostatných mojich názorov? Boli časy, keď ľudstvo nemalo takýto pojem. Niekto ho musel vymyslieť. Čo bolo to, čo si všimol? Všimol by som si, keby som videl niečo rovnako nové a rovnako dôležité? Dokázal by som myslieť tak ďaleko mimo vychodených koľají?

Koľko zo svojich vedomostí by ste dokázali obnoviť? Z akého hlbokého vymazania? To nie je iba test na vyhodenie nedostatočne prepojených názorov. Je to spôsob, ako vstrebať *fontánu poznania, nie iba jeden fakt.*

Bača postaví počítací systém, ktorý funguje tak, že vhodí kameň do vedra vždy keď ovca vyjde z ohrady a vyberie kameň vždy keď sa ovca vráti. Ak vy, ako učeň, nerozumiete tomuto systému – ak je to kúzlo, ktoré funguje bez zjavnej príčiny – potom nebudete vedieť, čo máte urobiť, ak vám náhodou do vedra padne jeden kameň navyše. To, čo nedokážete sami urobiť, nedokážete ani *prerobiť*, keď to situácia vyžaduje. Nemôžete sa vrátiť k zdroju, doladiť jeden z parametrov nastavenia a znovu vygenerovať výstup, ak nemáte zdroj. Ak je pre vás „Dva plus dva rovná sa štyri“ iba holý fakt a potom sa jeden z prvkov zmení na „päť“, ako zistíte, že „dva plus päť rovná sa sedem“, ak vám iba *povedali*, že „dva plus štyri rovná sa šesť“?

Ak nájdete rastlinu, z ktorej vypadne semienko vždy, keď nad ňou preletí vták, nenapadne vám, že by ste mohli túto rastlinu použiť na čiastočnú automatizáciu počítadla oviec. Napriek tomu, že ste sa naučili niečo, čo pôvodný autor mohol použiť na vylepšenie svojho vynálezu, vy sa neviete vrátiť k zdroju a znovu ho vytvoriť.

Keď obsiahnete zdroj myšlienky, samotná táto myšlienka sa môže meniť spolu s vami, ako budete nadobúdať nové vedomosti a nové zručnosti. Keď obsahujete zdroj myšlienky, stane sa naozaj vašou časťou a bude rásť spolu s vami.

60 Richard Rorty, „Out of the Matrix: How the Late Philosopher Donald Davidson Showed That Reality Can’t Be an Illusion,“ The Boston Globe (October 2003).

Snažte sa stať zdrojom každej myšlienky, ktorú je hodno myslieť. Ak táto myšlienka pôvodne prišla zvonku, postarajte sa, aby rovnako prichádzala aj zvnútra. Stále sa pýtajte: „Ako by som obnovil túto myšlienku, keby sa vymazala?“ Keď máte odpoveď, predstavte si, že by sa vymazala aj *táto* vedomosť. A keď nájdete fontánu, pozrite sa, čo iné z nej ešte vyteká.

⁂→

Medzihra: Jednoduchá pravda

> Pamätám si úlohu, ktorý som napísala o existencializme. Učiteľka mi ju vrátila s päťkou. Podčiarkla všetky slová pravda a pravdivý v eseji, bolo ich asi dvadsať, a vedľa každého napísala otáznik. Chcela vedieť, čo myslím slovom pravda.
>
> --Danielle Egan (novinárka)

Autorov predslov:

Cieľom tejto eseje je obnoviť naivný pohľad na pravdu.

Niekto vám povie: „Moja zázračná masť ťa dokáže zbaviť rakoviny pľúc za púhe tri týždne.“ Odpoviete: „Neukázala azda klinická štúdia, že toto tvrdenie nie je pravda?“ Dotyčný odvetí: „Takéto chápanie ‚pravdy‘ je veľmi naivné. Čo rozumieš slovom ‚pravda‘?“

Mnohí ľudia pri takejto otázke nevedia, ako dostatočne presne odpovedať. Napriek tomu by nebolo múdre zahodiť pojem „pravda“. Boli časy, keď nikto dostatočne presne nepoznal rovnice gravitácie, a predsa, kto skočil z útesu, ten spadol.

Často som videl – najmä v internetových diskusiách – ako uprostred rozhovoru niekto povie „X je pravda“ a potom sa začne debatovať o správnom používaní slova „pravda“. Táto esej *nie je* myslená ako encyklopedický odkaz pre takú debatu. Namiesto toho dúfam, že si debatujúci prečítajú túto esej a potom sa vrátia k tomu, o čom diskutovali predtým, než niekto spochybnil podstatu pravdy.

V tejto eseji kladiem otázky. Ak vidíte niečo, čo vyzerá ako celkom samozrejmá odpoveď, pravdepodobne je to odpoveď, ktorú som mal na mysli. Samozrejmá možnosť nemusí byť *vždy* tá najlepšia možnosť, avšak niekedy, šokujúco, *je*. Ak narazím na samozrejmú odpoveď, neprestanem hneď hľadať ďalej, ale ak hľadám ďalej a tá napohľad samozrejmá odpoveď *stále* vyzerá samozrejme, necítim sa previnilo, že sa jej držím. No iste, každý si *myslí*, že dva plus dva je štyri, každý *hovorí*, že dva plus dva je štyri, a aj v bežnej drine každodenného života sa každý *správa* ako keby dva plus dva bolo štyri, ale čomu sa dva plus dva *naozaj, v konečnom dôsledku* rovná? Nakoľko to dokážem posúdiť, je to štyri. Je to stále štyri, aj keď túto otázku vyrieknem vážnym, zlovestným tónom. Príliš jednoduché, hovoríte? Možnože v tomto prípade život *nemusí* byť zložitý. Nebolo by to osviežujúce?

Ak patríte k tým šťastným ľuďom, ktorým sa otázka zdá od začiatku jednoduchá, dúfam, že sa vám bude zdať jednoduchá aj na konci. Ak zistíte, že vás zarazili hlboké a významné otázky, pamätajte, že ak viete, ako presne systém funguje, a vedeli by ste si ho sami poskladať z vedier a kameňov, nemala by to pre vás byť záhada.

Ak vám hrozí zmätenie, keď interpretujete metaforu ako metaforu, skúste brať všetko *úplne doslova*.

* * *

Predstavme si, že som bača, v dobe pred zaznamenanou históriou formálnej matematiky a mám problém zrátať si ovce. Moje ovce spávajú v uzavretej ohrade, a tá ohrada je dosť vysoká na to, aby ovce chránila pred vlkmi blúdiacimi nocou. Každé ráno musím ovce vypustiť z ohrady na pašu; každý večer musím ovce nájsť a zavrátiť ich do ohrady. Ak nejaká ovca zostane vonku, ráno nájdem jej telo zabité a ohlodané vlkmi. Nechce sa mi celé hodiny prehľadávať lúky, hľadajúc poslednú stratenú ovcu, keď si myslím, že už sú asi všetky v ohrade. Niekedy to vzdám skôr a zvyčajne mi to prejde, ale asi každý desiaty raz nájde ráno ďalšiu mŕtvu ovcu.

→ http://lesswrong.com/lw/la/truly_part_of_you/

Kiež by existoval nejaký spôsob, ako zistiť, či sa ovce stále pasú, pohodlnejší než hľadanie! Skúšal som niekoľko metód: vrhal som veštecкými paličkami môjho kmeňa; cvičil som svoje duchovné schopnosti, aby som ovce našiel jasnovidectvom; starostlivo som vymenovával dôvody, prečo veriť, že všetky ovce sú v ohrade. Nič sa však nezmenilo. Asi každý desiaty raz, keď predčasne ukončím hľadanie, ráno nájdem mŕtvu ovcu. Môžem si uvedomovať, že moje metódy nefungujú, môžem starostlivo zdôvodňovať každú svoju chybu, ale dilema zostáva rovnaká. Buď strávim hodinu prehľadávaním každého možného zákutia, kde väčšinu času žiadna ovca nezostala; alebo idem spať skôr a prídem tak v priemere o jednu desatinu ovce.

Jedného večera som sa cítil zvlášť unavený. Vrhol som veštecкými paličkami a veštecké paličky povedali, že všetky ovce sa vrátili. Vizualizoval som si každé možné zákutie a nevybavila sa mi žiadna ovca. Stále som si nebol celkom istý, tak som sa pozrel do ohrady a zdalo sa, že je v nej veľa oviec, tak som si zrekapituloval svoje úsilie a rozhodol som sa, že som bol mimoriadne usilovný. To rozptýlilo moju úzkosť a šiel som spať. Ďalšie ráno som našiel *dve* mŕtve ovce. Čosi sa vo mne zlomilo a začal som uvažovať tvorivo.

V ten deň sa od brány košiara ozývali hlasné údery kladiva.

Nasledujúce ráno som bránu na ohrade iba odchýlil, a ako každá ovca vyšla z ohrady, vhodil som kamienok do vedra priklincovaného vedľa brány. Popoludní, ako každá ovca vošla, jeden kamienok som z vedierka vybral. Keď vo vedierku nezostal žiaden kameň, môžem prestať hľadať, a uložiť sa k spánku. Je to *skvelý* nápad. Bude to revolúcia bačovstva.

Toľko teória. V praxi bolo treba veľa vylepšení, než táto metóda začala fungovať spoľahlivo. Niekoľkokrát som celé hodiny hľadal a nenašiel žiadnu ovcu, a predsa nasledujúce ráno neboli žiadne straty. Každý z týchto prípadov si vyžadoval dôkladné premýšľanie, aby som zistil, kde môj vedrový systém zlyhal. Po návrate z jedného neplodného hľadania som sa zamyslel a uvedomil som si, že som ráno začal s vedrom, ktoré už nejaké kamene obsahovalo; zdá sa, že to bol zlý nápad. Inokedy som počas dňa náhodne hádzal kamienky do vedra, aby som zahnal nudu; to bol tiež zlý nápad, ako som si uvedomil po pár hodinách hľadania. Ale cvičil som sa v remesle kameňopočítania a stal som sa veľmi zručným kameňopočtárom.

Jedného popoludnia popri vyšliapanej ceste k mojim pastvinám kráčal bohato vyobliekaný muž v bielom rúchu, sandáloch a obleku.

„Môžem vám pomôcť“ pýtam sa.

Muž si vyberie z vrecka odznak a odklopí ho, čím nado všetku pochybnosť dokáže, že je Markos Sophisticus Maximus, poslanec Rumského senátu. (Človek by sa mohol zamyslieť, či sa taký odznak dá ukradnúť, ale moc týchto odznakov je taká veľká, že keby ich použil niekto iný, v okamihu by sa *premenil* na Markosa.)

„Volajte ma Mark,“ povie. „Prišiel som skonfiškovať čarovné kamienky v mene Senátu. Takéto mocné čarovné predmety sa nesmú dostať do rúk nevzdelancom.“

„Ten prekliaty učeň,“ zašomrem popod fúzy, „zase niečo natáral dedinčanom.“ Potom pozriem na Markovu prísnu tvár a vzdychnem si. „To nie sú čarovné kamienky,“ poviem nahlas. „Sú to obyčajné kamene, ktoré som zdvihol zo zeme.“

Markovo tvárou prebehne záblesk zmätenia, ale potom sa opäť rozžiari. „Prišiel som pre čarovné vedro!“ vyhlási.

„To nie je čarovné vedro,“ poviem unavene. „Predtým som doňho odkladal špinavé ponožky.“

Mark sa tvári zmätene. „Kde je teda to kúzlo?“ opýta sa.

Zaujímavá otázka. „To sa ťažko vysvetľuje,“ odpoviem.

Môj učeň Autrey, zaujatý udalosťou prichádza a ponúka svoje vysvetlenie: „Je to v hladine kamienkov vo vedre. Existuje čarovná hladina kamienkov, ktorú musíte mať celkom správne, inak to nefunguje. Ak vhodíte viac kamienkov, alebo nejaké vyberiete, vedro už nebude mať čarovnú hladinu. Práve teraz je čarovná hladina,“ Autrey nazrie do vedierka, „zhruba v jednej tretine.“

„Chápem!“ povie Mark vzrušene. Vytiahne z vrecka vlastné vedro a hromadu kamienkov. Vezme za hrsť kamienkov a vloží ich do vedra. Potom pozrie do vedra, koľko je tam kamienkov. „Aha,“ povie, „čarovná hladina tohto vedra je v polovici. Správne?“

„Nie!“ odsekne Autrey. „Hladina v polovici nie je čarovná. Čarovná hladina je zhruba v jednej tretine. Polovica je určite nečarovná. Navyše používate nesprávne vedro.“

Mark zmätene pozrie na mňa. „Povedali ste predsa, že vedro nie je čarovné?“

„Nie je,“ poviem. Okolo mňa vyjde z brány ovca, tak prihodím do vedra ďalší kameň. „Navyše, ja sledujem ovce. Rozprávajte sa s Autreyom.“

Mark s podozrením hľadí na kamienok, ktorý som práve prihodil, ale rozhodne sa odložiť túto otázku na neskôr. Otočí sa k Autreyovi a hrdo sa vystrie. „Toto je slobodná krajina,“ povie, „samozrejme, pod benevolentnou diktatúrou Senátu. Môžem ukladať hocijaké kamienky do hocijakého vedra.“

Autrey sa nad tým zamyslí. „Nie, nemôžete,“ povie nakoniec, „lebo v tom nebude kúzlo.“

„Pozrite,“ povie Mark trpezlivo. „Pozorne som vás sledoval. Pozreli ste sa do svojho vedra, skontrolovali ste hladinu kamienkov, a vyhlásili ste túto hladinu za čarovnú. Urobil som presne to isté.“

„Tak to nefunguje,“ povie Autrey.

„Aha,“ povie Mark. „Hladina kamienkov v *mojom* vedre nie je čarovná, ale vo *vašom* vedre áno. To tvrdíte? V čom je vaše vedro lepšie od môjho, há?“

Nuž, povie Autrey, *keby sme vyprázdnili vaše vedro, a potom presypali všetky kamienky z môjho vedra do vášho, potom by vaše vedro malo čarovnú hladinu. Existuje aj systém, ako skontrolovať, či má vaše vedro čarovnú hladinu, pokiaľ vieme, že ju má moje vedro; nazývame to operácia porovnania vedier.*

Prejde ďalšia ovca a ja prihodím ďalší kameň.

„Práve tam vhodil ďalší kameň!“ povie Mark. „Budete azda tvrdiť, že aj táto nová hladina je magická? Mohol by som do vášho vedra prihadzovať kamene, dokiaľ nebude jeho hladina rovnaká ako u mňa, a potom by sa naše vedrá zhodovali. Porovnávate moje vedro s vaším, aby ste určili, či je *podľa vás* hladina ‚čarovná‘ alebo nie. Nuž, z môjho pohľadu *vaše* vedro nie je čarovné, pretože nemá rovnakú hladinu kamienkov ako moje. Ha!“

„Moment,“ povie Autrey, „nechápete, že...“

„Slovom ‚čarovná hladina‘ jednoducho myslíte hladinu kamienkov vo svojom vlastnom vedre. A keď ja poviem ‚čarovná hladina‘, tak tým myslím počet kamienkov v mojom vedre. Takže keď vy pozriete na moje vedro a poviete, že nie je čarovné, slovo ‚čarovné‘ tu znamená rôzne veci pre rôznych ľudí. Musíte upresniť, *čia* čarovnosť to je. Môžete povedať, že moje vedro nemá ‚Autreyovu čarovnú hladinu‘ a ja poviem, že vaše vedro nemá ‚Markovu čarovnú hladinu‘. Týmto sa vyrieši domnelý paradox.“

„Ale...“ povie Autrey bezmocne.

„Rôzni ľudia môžu mať rôzne vedrá s rôznymi hladinami kamienkov, čo dokazuje, že celé vaše poňatie ‚čarovnosti‘ je naprosto subjektívne a svojvoľné.“

„Mark,“ poviem, „povedal vám niekto, na čo tie kamienky *slúžia*?“

„*Slúžia*?“ povie Mark. „Myslel som, že sú proste čarovné.“

„Keby tie kamienky nič nerobili,“ povie Autrey, „náš auditor efektivity procesov ISO 9000 by túto procedúru z našej dennej práce odstránil.“

„Kto je váš auditor?“

„Darwin,“ povie Autrey.

„Hm,“ povie Mark. „Charles má povesť prísneho auditora. Takže tieto kamienky požehnávajú stáda a zvyšujú tak počet oviec?“

„Nie,“ poviem. „Čaro kamienkov je v tomto: keď pozrieme do vedra a vidíme, že vedro je bez kamienkov, vieme, že aj pastviny sú bez oviec. Keď nepozrieme do vedra, musíme hľadať až do zotmenia, aby náhodou nejaká ovca nezostala vonku. Alebo ak skončíme predčasne, niekedy nájdeme na druhý deň mŕtvu ovcu, lebo vlci roztrhajú všetky zabudnuté ovce. Ak sa pozrieme do vedierka, vieme kedy sú všetky ovce doma, a môžeme ísť spať bez strachu.“

Mark nad tým uvažuje. „Znie to dosť nedôveryhodne," povie nakoniec. „Rozmýšľali ste nad použitím veštecký paličiek? Veštecké paličky sú neomylné; prinajmenšom každý, kto povie, že sú omylné, je upálený na hranici. A to je mimoriadne bolestivá smrť. Z toho vyplýva, že veštecké paličky sú neomylné."

„Vy môžete používať veštecké paličky, ak chcete," poviem.

„Ach, dobré nebesá, ani náhodou," povie Mark. „Paličky fungujú neomylne, absolútne dokonale v každom prípade, ako sa patrí na také požehnané nástroje. Ale čo keby na druhý deň bola nejaká ovca mŕtva? Ja používam veštecké paličky iba vtedy, keď neexistuje možnosť, že by sa ukázal ich omyl. Inak by ma mohli upáliť zaživa. Ako teda funguje vaše čarovné vedro?"

Ako funguje vedro...? Najlepšie bude začať tým najjednoduchším prípadom. „Asi takto," poviem, „predstavte si, že pastviny sú prázdne, ale vedro nie. Potom by sme strácali hodiny hľadaním ovce, ktorá tam nie je. A keby boli ovce na pastvinách, ale vedro by bolo prázdne, potom by sme Autrey a ja išli skôr spať, a na druhý deň by sme našli mŕtvu ovcu. Preto je prázdne vedro čarovné vždy vtedy a iba vtedy, keď sú pastviny prázdne..."

„Moment," povie Autrey. „To mi znie ako zbytočná tautológia. Nie sú snáď prázdne vedro a prázdne pastviny samozrejme tá istá vec?"

„Nie je to zbytočné," poviem. „Tu je analógia: Logik Alfred Tarski raz povedal, že veta ‚Sneh je biely' je pravdivá vždy vtedy a iba vtedy, keď sneh je biely. Ak rozumieš tomuto, potom by si mal chápať, prečo je prázdne vedro čarovné vždy vtedy a iba vtedy, keď sú pastviny bez oviec."

„Moment," povie Mark. „Toto sú *vedrá*. S *ovcami* nemajú nič spoločné. Vedrá a ovce sú samozrejme úplne odlišné. Neexistuje spôsob, ako by ovce mohli interagovať s vedrami."

„Odkiaľ teda podľa *vás* pochádza to kúzlo?" pýta sa Autrey.

Mark uvažuje. „Povedali ste, že môžete porovnať dve vedrá, aby ste zistili, či majú rovnakú hladinu... Chápem, ako môžu vedrá interagovať s vedrami. Možno keď máte veľké množstvo vedier a všetky majú rovnakú hladinu, *to* vytvára celé kúzlo. Budem to nazývať koherentistická teória čarovných vedier."

„Zaujímavé," povie Autrey. „Viem, že môj majster pracuje na systéme s viacerými vedrami – tvrdí, že môžu fungovať lepšie vďaka ‚redundancii' a ‚korekcii chýb'. To mi znie ako koherentizmus."

„To nie je to isté..." začnem namietať.

„Vyskúšajme si koherentistickú teória kúzla," povie Autrey. „Vidím, že máte vo vrecku ešte päť vedier. Podám vám vedro, ktoré používame, a vy potom naplníte vaše vedrá na rovnakú hladinu..."

Mark s hrôzou cúvne. „Stop! Tieto vedrá sa v mojej rodine odovzdávali celé generácie a vždy mali rovnakú hladinu! Ak prijmem vaše vedro, moja zbierka vedier sa stane menej koherentná a kúzlo sa stratí!"

„Ale vaše *terajšie* vedrá nemajú s ovcami nič spoločné!" namieta Autrey.

Mark vyzerá podráždene. „Pozrite, ako som už vysvetlil, samozrejme neexistuje spôsob, ako by ovce mohli interagovať s vedrami. Vedrá môžu interagovať iba s inými vedrami."

„Ja vhodím kameň vždy, keď vyjde ovca," podotknem.

„Keď vyjde ovca, vhodíte kameň?" povie Mark. „A ako to súvisí s našou témou?"

„Je to interakcia medzi ovcou a kameňmi," odpoviem.

„Nie, to je interakcia medzi kameňmi a *vami*," povie Mark. „Kúzlo nepochádza z oviec, pochádza z *vás*. Ovce samozrejme nie sú čarovné. Kúzlo musí *odniekadiaľ* pochádzať, aby sa dostalo do vedra."

Ukážem na drevený mechanizmus postavený na bráne. „Vidíte ten kus plachty, ktorá visí z tej drevenej hračky? Ešte stále sa s tým hráme – zatiaľ to nefunguje spoľahlivo – ale keď tadiaľ prejde ovca, potiahne plachtu. Keď sa plachta odhrnie, zo zásobníka vypadne kamienok a spadne do vedierka. Takto potom Autrey a ja nebudeme musieť vkladať kamienky osobne."

Mark zmraští obočie. „Nerozumiem tomu... je tá *plachta* čarovná?"

Pokrčím plecami. „Objednal som si ju cez internet od spoločnosti Prirodzený výber. Tá látka sa volá Zmyslová modalita." Zastavím sa, vidiac neveriace výrazy Marka a Autreya. „Pripúšťam, že tie

názvy znejú tak trochu ako new age. Pointa je, že prechod ovce spustí reťaz príčiny a účinku, ktorá končí kamienkom vo vedre. *Potom* už môžete porovnávať vedrá s vedrami, atď.“

„Tomu stále nerozumiem,“ povie Mark. „Nemôžete strčiť ovcu do vedra. Iba kamene idú do vedier, a je samozrejmé, že iba kamene môžu interagovať s inými kameňmi.“

„Ovce interagujú s vecami, ktoré interagujú s kameňmi...“ hľadám prirovnanie. „Predstavte si, že sa pozriete na svoje šnúrky na topánkach. Fotón vyletí zo Slnka, potom prejde zemskou atmosférou, potom sa odrazí od vašej šnúrky, potom prejde zreničkou vášho oka, potom dopadne na sietnicu, potom ho pohltí čapík alebo tyčinka. Energia fotónu spôsobí, že pripojený neurón vyšle signál, čo spôsobí, že ďalšie neuróny vyšlú signál. Vzor aktivácie neurónov vo vašej zrakovej kôre môže interagovať s vašimi myšlienkami o vašich šnúrkach, keďže vaše myšlienky o šnúrkach sú tiež uložené v neurónoch. Ak rozumiete tomuto, mali by ste chápať, ako prechádzajúca ovca spôsobuje, že kamienok padne do vedra.“

„A v *ktorom* presne okamihu počas tohto procesu sa kamienok stane čarovným?“ povie Mark.

„To... hm...“ Teraz začínam byť popletený *ja*. Potrasiem hlavou, aby som sa zbavil pavučín. Keď som sa dnes ráno zobudil, toto všetko mi pripadalo celkom jednoduché, a systém kamienkov a vedra sa medzičasom nestal o nič zložitejším. „Je to o dosť ľahšie na pochopenie, keď si pamätáte, že zmyslom celého systému je sledovať ovce.“

Mark smutne vzdychne. „No nič... je zrejmé, že neviete. Možno sú všetky kamienky čarovné už od začiatku, ešte predtým než sa dostanú do vedierka. Tento názor budeme nazývať pankamienkizmus.“

„Ha!“ povie Autrey s hlasom plným pohŕdania. „To sú len zbožné želania! Nie všetky kamienky boli stvorené ako rovné. Kamienky vo *vašom* vedre *nie* sú čarovné. Sú to len kusy kameňa!“

Mark sa zatvári prísne. „Teraz,“ vykríkne, „teraz vidíte, po akej nebezpečnej ceste kráčate! Keď raz poviete, že niektorí ľudia majú čarovné kamienky a iní nie, vaša pýcha vás pohltí! Budete sa považovať za lepších od druhých, a potom padnete na nos! Mnohí ľudia v dejinách mučili a vraždili, pretože si mysleli, že ich vlastné kamienky sú najlepšie!“ V Markovom hlase zaznie nádych blahosklonnosti. „Uctievať hladinu kamienkov ako ‚čarovnú‘ znamená, že existuje absolútna hladina v Najvyššom Vedre. Dnes už však nikto v Najvyššie Vedro neverí.“

„Po prvé,“ poviem, „ovce nie sú absolútne kamene. Po druhé, ja si nemyslím, že moje vedro naozaj obsahuje ovce. Po tretie, neuctievam hladinu svojho vedra ako čarovnú – občas ju zmením – a robím to *preto*, lebo mi záleží na ovciach.“

„Navyše,“ povie Autrey, „ak si niekto myslí, že mať absolútne kamienky by mu *dalo* právo mučiť a vraždiť, je to omyl, ktorý s vedrami nijako nesúvisí. Riešite tu nesprávny problém.“

Mark sa upokojí. „Predpokladám, že od púhych bačov nemožno očakávať nič viac. Vy zrejme veríte aj tomu, že sneh je biely, však?“

„Hmmm... áno?“ povie Autrey.

„A vôbec vám neprekáža, že aj *Stalin* veril, že sneh je biely?“

„Hmmm... nie?“ povie Autrey.

Mark sa neveriacky zahľadí na Autreya, a nakoniec pokrčí plecami. „Skúsme predpokladať, iba pre účely tejto diskusie, že vaše kamienky sú čarovné a moje nie sú. Viete mi povedať, v čom je ten rozdiel?“

„Moje kamienky *reprezentujú* ovce!“ povie Autrey víťazoslávne. „*Vaše* kamienky nemajú vlastnosť reprezentatívnosti, preto nefungujú. Chýba im zmysel. Len sa na ne pozrite. Nemajú auru sémantického obsahu, sú to púhe kamene. Potrebujete vedro so špeciálnou schopnosťou kauzality.“

„Ach!“ povie Mark. „Špeciálna schopnosť kauzality, namiesto kúzla.“

„Presne,“ povie Autrey. „Nie som poverčivý. Predpokladať kúzla, v dnešnej dobe, by bolo v medzinárodnej bačovskej komunite neprijateľné. Zistili sme, že predpokladať kúzla jednoducho nefunguje ako vysvetlenie pre bačovské fenomény. Takže keď vidím niečo, čomu nerozumiem, a chcem to vysvetliť pomocou modelu bez vnútorných detailov, ktorý neumožňuje žiadne predpovede, ani so spätnou platnosťou, použijeme predpoklad špeciálnej schopnosti. Ak nefunguje ani to, rozhodneme sa to nazývať emergentný fenomén.“

„Aké špeciálne schopnosti má toto vedro?“ pýta sa Mark.

„Hm,“ povie Autrey. „Možno je toto vedro nasiaknuté *vzťahovosťou* voči pastvinám. To by vysvetľovalo, prečo funguje – keď je vedro prázdne, *znamená* to, že pastviny sú prázdne.“

„Kde ste našli toto vedro?“ pýta sa Mark. „A ako ste prišli na to, že má vzťahovosť k pastvinám?“

„Je to *obyčajné vedro*,“ poviem. „Predtým som s ním liezol po stromoch... Nemyslím si, že to *treba* tak komplikovať.“

„Rozprávam sa s Autreyom,“ povie Mark.

„Vedro treba spojiť s pastvinami, a kamienky s ovcami, pomocou čarovného rituálu... prepáčte, pomocou emergentného procesu so špeciálnymi schopnosťami kauzality... ktoré môj majster objavil,“ vysvetlí Autrey.

Autrey sa potom pokúša opísať rituál, a Mark prikyvuje s múdrym porozumením.

„Musíte vhodiť kamienok *zakaždým*, keď cez bránu vyjde ovca?,“ pýta sa Mark. „Vybrať kamienok *zakaždým*, keď sa ovca vráti?“

Autrey prikyvuje. „Hej.“

„To musí byť veľmi ťažké,“ povie Mark súcitne.

Autrey sa rozžiari, pijúc Markov súcit ako dážď. „Presne tak!“ povie. „Je to *mimoriadne* emocionálne náročné. Keď malo vedro istý čas rovnakú hladinu, tak... si na tú hladinu zvyknete.“

Ovca vyjde z brány a prejde okolo nich. Autrey ju vidí, zastaví sa, dvihne kamienok a podrží ho vo vzduchu. „Ajhľa!“ vyhlási, „prešla ovca! Teraz musím vhodiť kameň do tohto vedra, moje drahé vedierko, a zničiť tým hladinu, ktorú si si tak dlho udržiavalo...“ Prejde ďalšia ovca. Autrey ju nezbadá, taký je zaujatý svojím vystúpením, takže jeden kameň prihodím ja. Autrey pokračuje: „...pretože to je najťažšou skúškou pravého bača, vhodiť kameň, akokoľvek to bolí, akokoľvek miloval pôvodnú hladinu. Iba tí najlepší spomedzi bačov splnia úlohu tak krutú...“

„Autrey,“ poviem, „ak raz chceš byť veľkým bačom, nauč sa držať hubu a hádzať kamienky. Žiaden rozruch. Žiadne divadlo. Proste to urob.“

„A tento rituál,“ povie Mark, „spája kamienky s ovcami podľa čarovných zákonov podobnosti a súvisu, ako bábika voodoo.“

Autrey sa mykne a obzrie. „Prosím vás! Nevolajte to podobnosť a súvis. My bačovia sme proti poverčivosti. Použite slovo ‚intencionalita‘ alebo niečo podobné.“

„Môžem sa pozrieť na ten kameň?“ opýta sa Mark.

„Iste,“ poviem. Vezmem jeden kamienok z vedierka a hodím ho Markovi. Potom sa zohnem, zdvihnem zo zeme ďalší kameň a vložím ho do vedra.

Autrey na mňa začudovane pozerá. „Nepokazil si to práve?“

Pokrčím plecami. „Myslím, že nie. Keby som to pokazil, zistíme to, ak zajtra nájdeme mŕtvu ovcu, alebo ak budeme celé hodiny hľadať a žiadnu ovcu nenájdeme.“

„Ale...“ povie Autrey.

„Naučil som ťa všetko, čo *ty* vieš, ale nie všetko, čo viem *ja*,“ poviem.

Mark skúma kamienok, sústredene naň hľadí. Drží nad ním svoju ruku a mrmle akési slová, potom potrasie hlavou. „Necítim žiadnu čarovnú moc,“ povie. „Prepáčte. Necítim žiadnu intencionalitu.“

„Kameň má intencionalitu iba vnútri čarovného, teda emergentného vedra,“ povie Autrey. „Inak je to púhy kameň.“

„To nie je problém,“ poviem. Vyberiem kamienok z vedierka a odhodím ho. Potom prídem k Markovi, potľapkám ho po ruke držiacej kamienok a poviem: „Vyhlasujem túto dlaň za súčasť čarovného vedra!“ Potom sa vrátim k bráne.

Autrey sa smeje: „Tak toto je od teba totálne zlomyseľné.“

Prikývnem, lebo je to naozaj tak.

„Ale bude to naozaj fungovať?“ pýta sa Autrey.

Opäť prikývnem, dúfajúc, že mám pravdu. Kedysi som to skúšal s dvoma vedierkami, a v princípe by nemal byť žiaden rozdiel medzi Markovou rukou a vedierkom. Dokonca aj keď Markova ruka

obsahuje *élan vital*, odlišujúci živú hmotu od neživej, trik by mal fungovať rovnako dobre, ako keby bol Mark mramorová socha.

Mark pozerá na svoju ruku, mierne nervózne. „Takže... teraz má kameň opäť intencionalitu?“

„Hej,“ poviem. „Neberte si do ruky viac kameňov, ani neodhadzujte tento jeden, inak porušíte rituál.“

Mark vážne prikývne. Potom pokračuje v skúmaní kamienka. „Teraz chápem, ako sa vaše stáda tak rozrastajú,“ povie. „Vďaka moci tohto vedra vám stačí vhadzovať kamene a ovce sa budú vracať z polí. Môžete začať iba s pár ovcami, vypustíte ich, a potom naplníte vedro až po okraj než sa vrátia. A keby vás starostlivosť o toľko ovce začala zaťažovať, môžete ich všetky vypustiť, potom vysypať takmer všetky kamene z vedra, takže sa vráti iba pár... opäť zvýšite ich počet, keď príde čas strihania... dobré nebesá, človeče! Uvedomujete si úžasnú *silu* tohto rituálu, ktorý ste objavili? Predstavte si jeho dôsledky; ľudstvo sa posunie vpred o desaťročie... nie, o storočie!“

„Takto to nefunguje,“ poviem. „Ak pridáte kameň, keď ovca neodišla, alebo odoberiete kameň, keď ovca neprišla, porušíte tým rituál. Moc nezostane v kameňoch, ale vyprchá, ako keď praskne mydlová bublina.“

Mark sa tvári hrozne sklamane. „Určite?“

Prikývnem. „Skúšal som to a nefungovalo to.“

Mark ťažko vzdychne. „Ach, táto... *matematika*... vyzerala taká mocná a užitočná, až kým... Nuž dobre. Zbohom, pokrok ľudstva.“

„Mark, ten nápad je *skvelý*,“ povie Autrey povzbudzujúco. „Mňa to vôbec nenapadlo a pritom je to také samozrejmé... ušetrilo by to *ohromné* množstvo námahy... *musí* existovať spôsob, ako váš plán zachrániť! Mohli by sme skúšať rôzne vedrá, hľadať také, ktoré uchová kúzlo... teda intencionalitu v kameňoch aj bez rituálu. Alebo vyskúšať iné kamene. Možno naše kamene jednoducho nemajú správne vlastnosti, a preto nemajú *vnútornú* intencionalitu. Čo keby sme skúšali použiť kamene vytesané do tvaru malých ovečiek? Alebo napísať na kamene slovo ‚ovca‘, to by mohlo stačiť.“

„To nebude fungovať,“ predpovedám sucho.

Autrey pokračuje. „Možno potrebujeme organické kamene namiesto kremíkových... alebo treba použiť drahé drahokamy. Ceny drahokamov sa zdvojnásobuje každých osemnásť mesiacov, takže keby ste teraz kúpili za hrsť lacných drahokamov a počkali, o dvadsať rokov by bolo naozaj drahé.“

„Vy ste skúšali pridávať kamene, aby ste vytvorili viac oviec, a nefungovalo to?“ pýta sa ma Mark. „Čo presne ste robili?“

„Zobral som za hrsť dolárových bankoviek. Potom som tie bankovky schoval pod perinu, jednu za druhou. Vždy keď som schoval ďalšiu bankovku, vzal som z krabičky jednu spinku na papier a uložil ju na kôpku. Dával som si pozor, aby som v hlave nedržal ich počet, takže som iba vedel, že mám ‚veľa‘ bankoviek a ‚veľa‘ spiniek. Potom, keď boli všetky bankovky ukryté pod perinou, pridal som na kôpku ešte jednu spinku na papier, čo je ekvivalentné prihodeniu jedného kamienka navyše do vedra. Potom som začal vyberať bankovky spod periny a ukladať spinky späť do krabičky. Keď som skončil, jedna spinka bola navyše.“

„Čo taký výsledok znamená?“ pýta sa Autrey.

„Znamená to, že trik nefungoval. Keď som porušil rituál tým jediným chybným krokom, moc nezostala, ale hneď vyprchala; kôpka spiniek a zbierka bankoviek sa už nemíňali v rovnakom čase.“

„Toto ste *naozaj* skúsili?“ pýta sa Mark.

„Áno,“ hovorím. „Naozaj som vykonal pokus, aby som overil, či výsledok zodpovedá mojim teoretickým predpovediam. Mám sentimentálny vzťah k vedeckej metóde, aj keď to vyzerá absurdne. Navyše, čo keby som sa mýlil?“

„Keby to *bolo* fungovalo,“ povie Mark, „boli by ste vinný z falšovania peňazí! Predstavte si, že by to robil každý; ekonomika by sa zrútila! Každý by mal miliardy dolárov, ale za tie peniaze by sa nedalo nič kúpiť!“

„Vôbec nie," odpoviem. „Podľa rovnakej logiky, kde pridanie ďalšej spinky na kôpku vytvorí ďalšiu dolárovú bankovku, vytvorenie ďalšej dolárovej bankovky by vytvorilo ďalší tovar a služby v hodnote jedného dolára."

Mark krúti hlavou. „Falšovanie peňazí je aj tak zločin... Nemali ste to skúšať."

„Bol som si *dosť* istý, že sa to nepodarí."

„Aha!" povie Mark. „*Očakávali* ste neúspech! *Neverili* ste, že to naozaj dokážete!"

„Veru tak", pripúšťam, „odhadli ste moje očakávania so zarážajúcou presnosťou."

„Nuž a v tom je problém", odvetí Mark rezko. „Kúzla sú poháňané vierou a silou vôle. Ak neveríte, že to dokážete urobiť, potom to nedokážete. Musíte zmeniť svoju vieru vo výsledok pokusu; tým sa zmení výsledok samotný."

„To je smiešne," poviem nostalgicky, „že presne toto mi Autrey povedal, keď som mu hovoril o metóde kameňa a vedra. Že preňho je úplne absurdné tomu uveriť, a preto to jemu fungovať nebude."

„Ako ste ho presvedčili?" vyzvedá Mark.

„Povedal som mu, aby držal hubu a riadil sa pokynmi," poviem, „a keď metóda fungovala, Autrey v ňu začal veriť."

Mark sa zmätene mračí. „To nedáva zmysel. Nerieši to základnú dilemu, či bolo skôr vajce alebo sliepka."

„Ale rieši. Metóda vedra funguje bez ohľadu na to, či v ňu veríte alebo nie."

„To je *absurdné*!" vyprskne Mark. „Neverím v existenciu kúziel, ktoré fungujú bez ohľadu na to, či v ne veríte alebo nie!"

„Aj ja som to hovoril," pridá sa Autrey. „Očividne som sa mýlil."

Mark sústredene mraští tvár. „Ale... ak ste neverili v kúzla, ktoré fungujú bez ohľadu na to, či v ne veríte alebo nie, prečo potom metóda vedra fungovala, keď ste v ňu neverili? Verili ste v kúzla, ktoré fungujú bez ohľadu na to, či v ne veríte alebo nie, bez ohľadu na to, či veríte v kúzla, ktoré fungujú bez ohľadu na to, či v ne veríte alebo nie?"

„Ja... *asi*... nie..." povedal Autrey neisto.

„Ak ste teda neverili v kúzla, ktoré fungujú bez ohľadu na to, či... vydržte chvíľku, musím si to nakresliť ceruzou na papier..." Mark fanaticky čmára, skepticky pozerá na výsledok, otáča papier hore nohami, potom sa vzdá. „Nechajte tak," povie. „Kúzla sú pre mňa dosť ťažké na pochopenie; metakúzla sú už na mňa priveľa."

„Mark, nemyslím si, že rozumiete umeniu kameňopočítania," hovorím. „To nie je o používaní kamienkov na ovládanie oviec. Ide o to, aby ovce ovládali kamienky. V tomto umení netreba začať tým, že veríte v jeho fungovanie. Naopak, toto umenie najprv funguje, a až potom začnete veriť, že funguje."

„Lebo tomuto veríte," povie Mark.

„Tomuto verím," odpoviem, „*pretože* je to naozaj tak. Súlad medzi skutočnosťou a mojimi názormi vzniká tak, že skutočnosť ovláda moje názory, nie naopak."

Prejde ďalšia ovca, čo spôsobí, že prihodím ďalší kamienok.

„Aha! Teraz sme sa dostali na koreň problému," povie Mark. „Čo je to tá takzvaná ‚skutočnosť'? Rozumiem, čo sa myslí tým, že nejaká hypotéza je elegantná, alebo falzifikovateľná, alebo v súlade s indíciami. Pripadá mi, že nazývať nejaký názor ‚pravdivý' alebo ‚skutočný' alebo ‚ozajstný', je jednoducho len rozdiel medzi povedaním, že si niečo myslíte, a že si niečo naozaj naozaj myslíte."

Zastavím sa. „No...." poviem pomaly. „Úprimne, tiež si nie som celkom istý, odkiaľ pochádza táto ‚skutočnosť'. Nedokážem si vytvoriť svoju vlastnú skutočnosť v laboratóriu, takže tomu zatiaľ dostatočne nerozumiem. Ale niekedy som silne presvedčený, že sa niečo stane, a potom sa stane niečo celkom iné. Potrebujem názov pre toto čosi, čo určuje výsledky mojich experimentov, takže to nazývam ‚skutočnosť'. Táto ‚skutočnosť' je čosi odlišné než moje hypotézy. Lebo aj keď mám jednoduchú hypotézu, v ktorej prospech svedčí mnoho mne známych indícií, aj tak ma niekedy výsledok prekvapí. Potrebujem teda rôzne názvy pre veci, ktoré určujú moje predpovede, a pre vec, ktorá určuje výsledky mojich experimentov. To prvé volám ‚názory' a to druhé volám ‚skutočnosť'."

Mark si odfrkne. „Ani neviem, prečo sa unúvam počúvať tieto jasné nezmysly. Čokoľvek poviete o tejto takzvanej ‚skutočnosti', je jednoducho ďalší názor. Dokonca aj váš názor, že najprv existuje skutočnosť a až potom naše názory, je názor. Z toho logicky nevyhnutne vyplýva, že skutočnosť neexistuje; existujú iba názory."

„Moment," hovorí Autrey, „mohli by ste mi zopakovať ten záver. Stratil som sa pri tej ostrej odbočke v strede."

„Bez ohľadu na to, čo hovoríte o skutočnosti, je to iba ďalší názor," vysvetľuje Mark. „Z toho totálne nevyhnutne vyplýva, že neexistuje žiadna skutočnosť, iba názory."

„Aha," poviem. „Takisto platí, že bez ohľadu na to, čo jeme, musíme to jesť ústami. Z toho vyplýva, že neexistuje žiadne jedlo, iba ústa."

„Presne tak," povie Mark. „Všetko, čo jete, musí byť vo vašich ústach. Ako by mohlo existovať nejaké jedlo mimo úst? Samotná táto myšlienka je nezmyselná, čo dokazuje, že ‚jedlo' je nekonzistentný pojem. Preto sme všetci na smrť hladní; žiadne jedlo neexistuje."

Autrey pozerá na svoje brucho. „Ale ja *nie som* na smrť hladný."

„Aha!" kričí Mark víťazoslávne. „A čím ste vyslovili samotnú túto námietku? Svojimi *ústami*, priateľu! *Ústami!* Aký lepší dôkaz neexistencie jedla by ste si mohli žiadať?"

„*Čo to tu počujem o hladovaní?*" opýta sa drsný chrapľavý hlas priamo spoza nás. Autrey a ja zostaneme v pohode, už sme si týmto prešli. Mark vyplašene vyskočí pol metra do vzduchu.

Inšpektor Darwin sa usmeje zovretými perami, potešený z úspešného prekvapenia, a spraví si malú čiarku vo svojom poznámkovom bloku.

„To je len metafora!" povie Mark rýchlo. „Nemusíte mi kvôli tomu brať ústa, ani nič také..."

„*Načo* sú vám *ústa*, ak neexistuje žiadne *jedlo*?" pýta sa Darwin zlostne. „*No nič*. Nemám *čas* na takéto *hlúposti*. Prišiel som na inšpekciu *oviec*."

„Stádam sa darí, pane," hovorím. „Od januára nezomrela žiadna ovca."

„*Výborne*. Prideľujem vám 0,12 bodu *úspechu*. A teraz, čo tu robí táto *osoba*? Je nevyhnutnou súčasťou tejto *operácie*?"

„Zatiaľ sa mi zdá, že by bol ľudstvu užitočnejší, keby ho zavesili z balóna ako závažie," hovorím.

„Ajaj," povie Autrey potichu.

„Úžitok *ľudstva* ma vôbec *nezaujíma*. Nech sa vyjadrí *sám*."

Mark sa rýchlo narovná. „Tento púhy *bača*," povie, ukazujúc na mňa, „tvrdí, že existuje niečo také ako skutočnosť. To ma uráža, pretože ja viem s hlbokým a trvalým presvedčením, že žiadna pravda neexistuje. Koncept ‚pravdy' je iba strategický manéver ľudí, ktorí chcú vnucovať svoje názory druhým. Každý kultúra má inú ‚pravdu' a ‚pravda' žiadnej kultúry nie je nadradená iným. Čo som práve povedal, platí v každom čase a na každom mieste, a trvám na tom, aby ste súhlasili."

„Počkajte chvíľu," hovorí Autrey. „Ak nič nie je pravda, prečo by som mal veriť tomu, keď hovoríte, že nič nie je pravda?"

„Nepovedal som, že nič nie je pravda..." hovorí Mark.

„Ale áno," preruší ho Autrey, „počul som vás."

„...povedal som, že ‚pravda' je výhovorka používaná niektorými kultúrami na vnucovanie svojich názorov druhým. Keď poviete, že niečo je ‚pravda', myslíte tým len to, že by pre vašu spoločenskú skupinu bolo výhodné, keby sa tomu verilo."

„A toto, čo ste práve povedali," hovorím, „to je pravda?"

„Absolútna a jednoznačná pravda!" povie Mark precítene. „Ľudia si tvoria svoje vlastné skutočnosti."

„Počkajte," povie Autrey, opäť znejúci zmätene, „povedať, že ľudia si vytvárajú svoje vlastné skutočnosti, je logicky úplne iná vec ako povedať, že neexistuje žiadna pravda, čo je stav, ktorý si ani neviem koherentne predstaviť, azda pretože ste stále nevysvetlili, ako presne to funguje..."

„A zase to tu máme," povie Mark podráždene, „pokúšate sa aplikovať svoj západný koncept logiky, rozumnosti, rozumu, koherencie a vnútornej konzistencie."

„No super," šomre Autrey, „tak sme zároveň načali ešte *tretiu*, úplne nesúvisiacu a odlišnú tému..."

„Nie je nesúvisiaca,“ hovorí Mark. „Pozrite, beriete to zo zlého uhla, ak chápete moje výroky ako hypotézy a snažíte sa odvodiť ich dôsledky. Mali by ste ich brať ako úplne všeobecné výhovorky, ktoré používam, keď niekto povie niečo, čo sa mi nepáči. Nemá to byť model toho, ako funguje vesmír, ale skôr ako karta ‚môžete slobodne vyjsť z väzenia‘. *Kľúčom* je uplatňovať tieto výhovorky *selektívne*. Keď poviem, že neexistuje nič také ako pravda, uplatňujem to iba na *vaše* tvrdenie, že čarovné vedierko funguje bez ohľadu na to, či v to verím alebo nie. *Neuplatňujem* to na *svoje* tvrdenie, že neexistuje nič také ako pravda.“

„Hmmm... prečo nie?“ vyzvedá Autrey.

Mark si unavene vzdychne. „Autrey, myslíte si, že ste prvý človek, ktorý kladie túto otázku? Ktorý sa pýta, ako naše vlastné názory môžu mať zmysel, keď žiadne názory nemajú zmysel? Tú istú vec sa pýtajú mnohí študenti, keď sa stretnú s touto filozofiou, o ktorej vás musím informovať, že má veľa priaznivcov a rozsiahlu literatúru.“

„Aká je teda odpoveď?“ pýta sa Autrey.

„Nazývame to ‚problém reflektivity‘,“ vysvetlí Mark.

„Ale aká je *odpoveď*?“ trvá na svojom Autrey.

Mark sa povznesene usmieva. „Verte mi, Autrey, nie ste prvý človek, ktorého napadla takáto jednoduchá otázka. Nemá zmysel predkladať nám ju ako nejaké triumfálne vyvrátenie.“

„Ale aká je tá *skutočná odpoveď*,?“

„Teraz by som rád presunul k téme, ako logika zabíja malé tulenie mláďatká...“

„*Zabíjate čas*,“ vyštekne inšpektor Darwin.

„Navyše, nesledujete ovce,“ poviem a prihodím ďalší kamienok.

Inšpektor Darwin pozerá na dvoch diskutérov, očividne neochotných ustúpiť zo svojich pozícií. „Počujte,“ povie Darwin, tentokrát prívetivejšie. „Mám jednoduchý návrh, ako vyriešiť vašu dišputu. *Vy* tvrdíte,“ povie, ukazujúc na Marka, „že názory ľudí menia ich osobnú skutočnosť. A *vy* pevne veríte,“ jeho prst sa otočí smerom na Autreya, „že Markove názory *nedokážu* zmeniť skutočnosť. Takže nechajme Marka naozaj pevne veriť, že vie lietať, a nech skočí z tohto útesu. Mark uvidí sám seba odletieť ako vták a Autrey ho uvidí padnúť dole a rozpleštiť sa, obaja budete spokojní.“

Zastavíme sa a uvažujeme o tom.

„*Znie* to rozumne...“ povie Mark nakoniec.

„Útes je hneď tu,“ konštatuje inšpektor Darwin.

Autrey sa tvári mimoriadne sústredene. Nakoniec vykríkne: „Počkajte! Keby to bola pravda, my všetci by sme sa už dávno odseparovali každý do svojho vlastného vesmíru, a v tom prípade by všetci ostatní ľudia boli iba výtvormi našej predstavivosti... a potom nemá zmysel nám niečo dokazovať...“

Z blízkeho útesu zaznie dlhý, postupne tíchnuci výkrik, nasledovaný tupým šplechnutím a tichom. Inšpektor Darwin vo svojom zápisníku nalistuje stranu zobrazujúcu súčasný genofond a zapíše si o čosi nižšiu frekvenciu pre Markove alely.

Autrey vyzerá mierne zdesene. „Bolo to naozaj nutné?“

„*Nutné?*“ povie inšpektor Darwin nechápavo. „Proste sa to *stalo*... Nerozumiem celkom vašej otázke.“

Autrey a ja sa vrátime k nášmu vedierku. Je čas zaháňať ovce. Na túto časť nesmiete zabudnúť. Pretože inak, aký by to celé malo zmysel?

*→

→ http://yudkowsky.net/rational/the-simple-truth

Kniha II.

Ako naozaj zmeniť svoj názor

Rozumnosť: Úvod

(napísal Rob Bensinger)

> Čo by som si mal myslieť?
> Ako sa ukazuje, táto otázka má správnu odpoveď.

Má správnu odpoveď, aj keď vás trápi neistota, nielen keď máte definitívny dôkaz. Vždy existuje správne množstvo istoty, ktoré by ste mali dať tvrdeniu, dokonca aj keď vyzerá ako „osobný názor“ a nie ako „fakt“ overený odborníkmi.

Napriek tomu často hovoríme akoby existencia neistoty a nesúhlasu robila z názorov púhu otázku vkusu. Hovoríme „toto je len môj názor“ alebo „máš právo na vlastný názor“ akoby tvrdenia fyziky a matematiky existovali v nejakej inej, vyššej rovine ako názory, ktoré sú iba „súkromné“ a „subjektívne“. Avšak, Robin Hanson píše:[61]

> Nemáte právo na vlastný názor. Nikdy! Nemáte ani len právo na „neviem“. Máte právo na svoje túžby, a niekedy na svoje rozhodnutia. Môžete si privlastniť rozhodnutie, ak si dokážete vybrať svoje preferencie, možno máte právo tak urobiť. Ale vaše názory nie sú o vás; názory sú o svete. Vaše názory by mali byť vašimi najlepšími odhadmi toho, ako veci sú; čokoľvek iné je lož. (…)
>
> Je pravda, že niektoré témy dávajú odborníkom silnejšie mechanizmy na riešenie sporov. Pri iných témach naše skreslenia a zložitosť sveta robia vytváranie silných záverov ťažším. (…)
>
> Nikdy však nezabudnite, že na každú otázku ohľadom toho, ako veci sú (alebo by mali byť), a v ľubovoľnej informačnej situácii, vždy *existuje* nejaký najlepší odhad. Máte akurát právo na vlastné najlepšie poctivé úsilie nájsť tento najlepší odhad; čokoľvek iné je lož.

Predstavte si, že zistíte, že jeden zo šiestich ľudí je do vás zamilovaný – možno dostanete list od tajného obdivovateľa a ste si istý, že je od jedného z týchto šiestich – ale netušíte, ktorý z tých šiestich to je. Váš spolužiak Bob je jedným z týchto šiestich kandidátov, ale nemáte žiadne špeciálne indície v prospech ani neprospech toho, že on je ten zamilovaný. V tom prípade je šanca, že Bob je tým zamilovaným, 1 : 5.

Pretože existuje šesť možností, náhodný tip by spôsobil, že v priemere uhádnete raz správne na každých päť prípadov, keď ste hádali nesprávne. To je to, čo myslíme slovami „šanca je 1 : 5“. Nemôžete povedať: „No, ja netuším, kto je do mňa zamilovaný; možno je to Bob, možno nie je. Takže poviem, že šanca je päťdesiat na päťdesiat.“ Aj keby ste radšej povedali „neviem“ alebo „možno“ a tým skončili, odpoveď je stále 1 : 5.[62]

Predpokladajme, že ste si ďalej všimli, že keď sú ľudia do vás zamilovaní, žmurkajú na vás desaťkrát častejšie. Ak vás Bob zažmurká, je to nový kus indície. V tomto prípade by bolo chybou zostať skeptickým ohľadom toho, či je Bob váš tajný obdivovateľ; šanca 10 : 1 v prospech „náhodná osoba, ktorá na mňa žmurká, je zamilovaná“ prevažuje šancu 1 : 5 proti „Bob je do mňa zamilovaný“.

Bolo by *tiež* chybou povedať: „Táto indícia je taká silná, že je to tutovka, že on je ten, kto je do mňa zamilovaný! Odteraz budem skrátka predpokladať, že Bob je do mňa.“ Prehnaná istota je rovnako zlá ako nedostatočná.

V skutočnosti na túto otázku existuje iba jedna matematicky konzistentná odpoveď. Aby sme zmenili svoju myseľ z pôvodnej šance 1 : 5 podľa pomeru podmienených pravdepodobností 10 : 1, vynásobíme dokopy ľavé strany a dokopy pravé strany, čím dostaneme výslednú šancu 10 : 5, čiže 2 : 1 v prospech „Bob je do mňa zamilovaný“. Pri daných predpokladoch a dostupných indíciách, hádanie, že Bob je do vás zamilovaný, sa ukáže ako správne 2-krát na každý 1 raz, keď sa ukázalo ako nesprávne. Ekvivalentne: pravdepodobnosť, že ho priťahujete, je 2/3. Ľubovoľná iná úroveň istoty by bola nekonzistentná.

Naša kultúra si nezvnútornila lekcie teórie pravdepodobnosti – že správna odpoveď na otázky ako: „Nakoľko si môžem byť istá, že je do mňa Bob zamilovaný?“ je rovnako logicky určená ako správna

---

61 Robin Hanson, „You Are Never Entitled to Your Opinion,“ *Overcoming Bias (blog)* (2006), http://www.overcomingbias.com/2006/12/you_are_never_e.html.

62 Toto vyplýva z predpokladu, že existuje šesť možností, a že nemáte žiaden dôvod uprednostňovať niektorú z nich pred hociktorou z ostatných. Takisto predpokladáme, nerealisticky, že si môžete byť naozaj istý, že ten obdivovateľ je jeden z týchto šiestich ľudí, a že nezanedbávate iné možnosti. (Čo ak je do vás zamilovaných viac ako jeden z tých šiestich?)

odpoveď na otázku v písomke z matematiky alebo v učebnici zemepisu. Naše klišé zaostávajú za objavom, že „aké názory by som mal mať?“ má objektívne správnu odpoveď, či už je vašou otázkou „je do mňa môj spolužiak zamilovaný?“ alebo „mám nesmrteľnú dušu?“ Naozaj existuje správny spôsob, ako zmeniť svoj názor. A je to *presný* spôsob.

Ako v skutočnosti nezmeniť svoj názor

Upravovať svoje názory na hocičo spôsobom podobným tomuto idealizovanému postupu je však zapeklitá úloha.

V prvej časti knihy *Rozumnosť: od algoritmov po zombie* sme diskutovali o hodnote „skutočných“ názorov. Nie je v princípe nič zlé na vyjadrovaní podpory niečomu, na čom vám záleží – napríklad skupine, s ktorou sa identifikujete, alebo duchovnému zážitku, ktorý považujete za povznášajúci. Keď si však mýlime slogany s faktickými názormi, dokážu tieto nepochopené slogany zaštítiť celú ideológiu pred poškvrnením indíciami.

Ešte ani názory, ktoré vyzerajú, že elegantne vysvetľujú naše pozorovania, nie sú imúnne voči tomuto problému. Je pre nás príliš ľahké vidieť hmlistú vedecky znejúcu (alebo inak autoritatívnu) vetu a dôjsť k záveru, že niečo „vysvetlila“, hoci neovplyvnila šance, ktoré implicitne priraďujeme našim možným budúcim zážitkom.

Čo je najhoršie, prozaické názory – názory, ktoré *sú* v princípe falzifikovateľné, názory, ktoré *obmedzujú* to, čo očakávame, že uvidíme – sa nám stále môžu zaseknúť v hlavách, posilnené sieťou ilúzií a skreslení.

V roku 1951 sa futbalový zápas medzi Dartmouthom a Princetonom vyvinul nezvyčajne drsne. Psychológovia Hastorf a Cantril sa pýtali študentov z každej školy, kto začal hrať drsne. Takmer všetci sa zhodli na tom, že nezačal Princeton; ale 86 % študentov Princetonu verilo, že to začal Dartmouth, zatiaľ čo iba 36 % študentov Dartmouthu obviňovali Dartmouth. (Väčšina študentov Dartmouthu verila, že „začali obe strany“.)

Nie je dôvod myslieť si, že to bol slogan a nie skutočný názor. Študentov pravdepodobne viedli ich rôzne názory k rôznym predpovediam o správaní hráčov v budúcich zápasoch. A predsa akosi boli dokonale obyčajné faktické názory v Dartmouthe výrazne odlišné od dokonale obyčajných faktických názorov v Princetone.

Môžeme za toto viniť rôzne zdroje, ku ktorým mali prístup študenti Dartmouthu a Princetonu? Skreslenie v rôznych zdrojoch novín, na ktoré sa rôzne skupiny spoliehajú, je sám osebe celkom vážny problém.

Avšak v tomto prípade tu pôsobí ešte niečo iné. Keď študentom neskôr naozaj *ukázali* filmový záznam zápasu a požiadali ich, aby spočítali priestupky, ktoré vidia, študenti Dartmouthu tvrdili, že vidia v priemere 4,3 priestupku zo strany tímu Dartmouthu (a polovicu z nich klasifikovali ako „mierne“), zatiaľ čo študenti Princetonu tvrdili, že vidia v priemere 9,8 priestupku zo strany tímu Dartmouthu (a tretinu z nich klasifikovali ako „mierne“).

Zabudnime na to, že by sa súperiace frakcie dokázali zhodnúť ohľadom zložitých návrhov v štátnej politike alebo v morálnej filozofii; študenti verní rôznym skupinám sa nedokázali zhodnúť ani na tom, čo *videli*.[63]

Keď je ohrozené niečo, na čom nám záleží – náš svetonázor, členovia našej skupiny, naše spoločenské postavenie, alebo hocičo iné – naše myšlienky a vnemy tomu utekajú na pomoc.[64,65] Niektorí

63 Albert Hastorf and Hadley Cantril, „They Saw a Game: A Case Study,“ *Journal of Abnormal and Social Psychology* 49 (1954): 129–134, http://www2.psych.ubc.ca/~schaller/Psyc590Readings/Hastorf1954.pdf.

64 Pronin, „How We See Ourselves and How We See Others.“

65 Robert P. Vallone, Lee Ross, and Mark R. Lepper, „The Hostile Media Phenomenon: Biased Perception and Perceptions of Media Bias in Coverage of the Beirut Massacre,“ *Journal of Personality and Social Psychology* 49 (1985): 577–585, http://ssc.wisc.edu/~jpiliavi/965/hwang.pdf.

psychológovia dnes idú až tak ďaleko, že tvrdia, že naša schopnosť nachádzať slovné zdôvodnenia pre naše závery sa vyvinula *konkrétne* preto, aby nám pomáhala vyhrať hádky.[66]

Jeden z charakteristických vhľadov psychológie 20. storočia, ktorý motivuje každého od Freudových žiakov po súčasných kognitívnych psychológov je, že ľudské správanie často poháňajú sofistikované nevedomé procesy, a že príbehy, ktoré sami sebe hovoríme o svojich motívoch a dôvodoch sú omnoho skreslenejšie a vymyslenejšie než si uvedomujeme.

Často sa nám vlastne ani nedarí uvedomiť si, že práve rozprávame nejakú rozprávku. Keď sa nám v rámci sebapozorovania zdá, že „priamo vnímame" veci o sebe, často sa ukáže, že sa zakladajú na nespoľahlivých implicitných kauzálnych modeloch.[67,68] Keď sa pokúšame argumentovať v prospech svojich názorov, dokážeme si vymyslieť pochybné dôvody, ktoré nemajú nič spoločné s tým, ako sme k tomuto názoru po prvýkrát došli.[69] Namiesto posudzovania našich vysvetlení podľa ich schopnosti predpovedať, rozprávame príbehy, aby sme dali zmysel tomu, čo si myslíme, že vieme.

Ako to môžeme robiť lepšie? Ako môžeme dospieť k realistickému pohľadu na svet, keď naše mysle majú taký sklon racionalizovať? Ako môžeme dospieť k realistickému pohľadu na svoj myšlienkový život, keď naše myšlienky *o myslení* sú tiež podozrivé? Ako sa môžeme stať menej skreslenými, keď samotné naše snahy zbaviť sa skreslení môžu mať svoje vlastné skreslenia?

Čo je *najmenej* pochybné miesto, o ktoré by sme sa mohli oprieť?

Matematika rozumnosti

Na prelome 20. storočia vymyslenie jednoduchých axióm aritmetiky (napríklad založených na teórii množín) dalo matematikom jasnejší štandard na posudzovanie správnosti ich záverov. Ak človek alebo kalkulačka vypíše „2 + 2 = 4", môžeme dnes urobiť viac než len povedať: „zdá sa mi to intuitívne správne". Môžeme vysvetliť, *prečo* to je správne, a môžeme dokázať, že jeho správnosť systematickým spôsobom súvisí so správnosťou zvyšku aritmetiky.

Lenže matematika a logika nám dovoľujú modelovať správanie fyzikálnych systémov, ktoré sú omnoho zaujímavejšie než vrecková kalkulačka. Môžeme formalizovať aj *rozumný názor vo všeobecnosti* pomocou teórie pravdepodobnosti, aby sme vyzdvihli črty, ktoré majú spoločné všetky úspešné formy usudzovania. Môžeme dokonca formalizovať aj *rozumné správanie vo všeobecnosti* pomocou teórie rozhodovania.

Teória pravdepodobnosti definuje, ako by sme ideálne rozmýšľali zoči-voči neistote, keby sme mali potrebný čas, výpočtovú kapacitu, a sebaovládanie. Pri danom doterajšom poznaní (pôvodné predpoklady) a novom kuse indície, teória pravdepodobnosti jednoznačne definuje najlepšiu množinu nových názorov (výsledok), ktoré by som mal prijať. Podobne teória rozhodovania definuje, aké činy by som mal urobiť na základe svojich názorov. Pre ľubovoľnú konzistentnú množinu názorov a preferencií, aké by som mohol mať ohľadom Boba, teória rozhodovania dáva odpoveď, ako by som potom mal konať, aby som uspokojil svoje preferencie.

Ľudia nedokážu dokonale rozmýšľať ani sa dokonale rozhodovať, rovnako ako nie sme dokonalé kalkulačky. Naše mozgy sú zlepencom, ktorý pozliepal dokopy prirodzený výber. Pri najlepšej vôli nedokážeme vypočítať *presne* správnu odpoveď na „čo by som si mal myslieť?" a *čo by som mal robiť?* Chýba nám na to čas a výpočtová kapacita, a evolúcia nemala inžiniersku odbornosť a predvídavosť, aby vyžehlila všetky naše chyby.

66 Hugo Mercier and Dan Sperber, „Why Do Humans Reason? Arguments for an Argumentative Theory," *Behavioral and Brain Sciences* 34 (2011): 57–74, https://hal.archives-ouvertes.fr/file/index/docid/904097/filename/-MercierSperberWhydohumansreason.pdf.

67 Richard E. Nisbett and Timothy D. Wilson, „Telling More than We Can Know: Verbal Reports on Mental Processes," *Psychological Review* 84 (1977): 231–259, http://people.virginia.edu/~tdw/nisbett&wilson.pdf.

68 Eric Schwitzgebel, *Perplexities of Consciousness* (MIT Press, 2011).

69 Jonathan Haidt, „The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment," *Psychological Review* 108, no. 4 (2001): 814–834, doi:10.1037/0033-295X.108.4.814.

Maximálne efektívny bezchybný uvažovateľ v skutočnom svete by sa v skutočnosti stále potreboval opierať o heuristiky a približné výpočty. Optimálne výpočtovo zvládnuteľné algoritmy na zmenu názorov stále nestačia na konzistenciu teórie pravdepodobnosti.

Ale predsa, hoci vieme, že sa nemôžeme stať *celkom* konzistentnými, stále sa určite môžeme zlepšiť. Vedomie, že existuje ideálny štandard, s ktorým sa môžeme porovnávať – ktorý výskumníci nazývajú „bayesovská rozumnosť“ - nás môže viesť k zdokonaľovaniu našich myšlienok a činov. Hoci nikdy nebudeme dokonalí Bayesovci, matematika rozumnosti nám môže pomôcť pochopiť, *prečo* je nejaká odpoveď správna, a pomôcť nám zbadať presné miesto, kde sme sa pomýlili.

Predstavte si, že sa pokúšate naučiť matematiku výlučne pomocou memorovania. Môžu vám povedať, že „10 + 3 = 13“ a „31 + 108 = 139“ atď., ale príliš vám to nepomôže, dokiaľ nepochopíte systém za týmito znakmi. Môže byť omnoho ťažšie hľadať metódy na zlepšenie svojej rozumnosti, keď nemáte všeobecný rámec na hodnotenie úspechu danej metódy. Cieľom tejto knihy je pomôcť ľuďom vybudovať si pre seba takéto rámce.

Aplikovaná rozumnosť

V článku na blogu pojednávajúcom o tom, ako sa odlišujú rozumnosťou nadšení „racionalisti“ od anti-empirických „racionalistov“, Scott Alexander hovorí:[70]

> Samozrejme je užitočné mať toľko indície, koľko sa len dá, rovnako ako je užitočné mať toľko veľa peňazí, ako sa len dá. Ale rovnako samozrejme je užitočné vedieť múdro využiť obmedzené množstvo indície, rovnako ako je užitočné vedieť múdro využiť obmedzené množstvo peňazí.

Techniky rozumnosti nám pomáhajú vyťažiť viac z indícií, ktoré máme, v prípadoch, kde táto indícia nie je dostatočná, alebo kde naše skreslenia a pripútanosť krivia spôsob, akým tieto indície interpretujeme. To sa týka našich osobných životov, ako v prípade Boba. Týka sa to nesúhlasov medzi politickými frakciami (a medzi športovými fanúšikmi). A týka sa to technologických a filozofických hlavolamov, ako je debata o transhumanizme, o postoji, že by sme mali použiť technológiu na dôkladné renovovanie stavu človeka. Uvedomiť si, že rovnaké matematické zákony platia v každej z týchto oblastí – a že v mnohých prípadoch vládnu rovnaké kognitívne skreslenia - *Ako naozaj zmeniť svoj názor* čerpá zo širokej palety ukážkových príkladov.

Prvá postupnosť esejí v *Ako naozaj zmeniť svoj názor*, „Príliš pohodlné výhovorky“, sa sústredí na otázky, ktoré sú pravdepodobnostne také jasné, aké len môžu byť. Bayesovsky optimálnu odpoveď je často nereálne vypočítať, ale skreslenia ako sklon potvrdzovať sa dokážu zakoreniť aj v prípadoch, kde je dostupnej indície vyše hlavy a máme dostatok času veci si premyslieť.

Odtiaľ sa pohneme do mútnejších vôd v postupnosti „Politika a rozumnosť“. Klasická štátna politika, ako ju debatujú televízni experti, je známa svojimi zlostnými, neproduktívnymi diskusiami. Keď sa na to pozrieme, je na tom čosi prekvapivé. Prečo berieme politický nesúhlas tak osobne, dokonca aj keď sú mechanizmus a účinky štátnej politiky od nás tak ďaleko v priestore alebo v čase? A keď už sme pritom, prečo sa nestávame *viac* opatrní a presní s indíciami, keď sa jedná o témy, ktoré považujeme za dôležité?

Zápas Dartmouth-Princeton naznačuje odpoveď. Mnohé z nášho procesu uvažovania je v skutočnosti racionalizácia – rozprávanie príbehov, ktoré dávajú našim terajším názorom pocit väčšej súdržnosti a zdôvodnenia, hoci nevyhnutne nezlepšujú ich presnosť. „Proti racionalizácii“ hovorí o tomto probléme, a nasleduje „Proti doublethinku“ (o sebaklame) a „Videnie čerstvými očami“ (o náročnosti všimnutia si indície, ktorá nezodpovedá našim očakávaniam a predpokladom).

Postúpiť na vyššiu úroveň rozumnosti znamená stretnúť sa s mnohými zaujímavými a mocnými novými myšlienkami. V mnohých prípadoch to znamená aj robiť si priateľov, s ktorými si môžete vymieňať myšlienky a nachádzať spoločenstvá, ktoré vás povzbudzujú, aby ste to robili lepšie. „Špirály

70 Scott Alexander, „Why I Am Not Rene Descartes,“ *Slate Star Codex (blog)* (2014), http://slatestarcodex.com/-2014/11/27/why-i-am-not-rene-descartes/.

smrti“ rozoberá niektoré dôležité nebezpečenstvá, ktoré môžu postihnúť skupiny spojené spoločnými záujmami a úžasnými žiarivými myšlienkami, ktoré bude treba prekonať, ak chceme dostať plný úžitok z racionalistických spoločenstiev. *Ako naozaj zmeniť svoj názor* potom končí postupnosťou „Zanechávanie“.

Naším prirodzeným stavom *nie je* meniť svoje názory tak, ako by to robil Bayesovec. Primäť študentov Dartmouthu a Princetonu, aby si *všimli, čo naozaj vidia*, nebude také jednoduché ako odrecitovať im axiómy teórie pravdepodobnosti. Ako píše Luke Muehlhauser v Sile konania:[71]

> Vy nie ste bayesovský škriatok, ktorého uvažovanie „kazia“ kognitívne skreslenia.
>
> Vy jednoducho *ste* kognitívne skreslenia.

Sklon potvrdzovať, sklon k status quo, sklon prisudzovať a podobné veci nie sú prilepené k nášmu rozmýšľaniu; oni sú jeho samotnou podstatou.

To neznamená, že zbaviť sa skreslení je nemožné. Takisto nie sme ani dokonalé kalkulačky pod všetkými svojimi aritmetickými chybami. Mnohé z našich matematických obmedzení vyplývajú z veľmi hlbokých faktov o tom, ako funguje ľudský mozog. Ale predsa si dokážeme nacvičiť matematické schopnosti; dokážeme sa naučiť, kedy dôverovať a kedy nedôverovať svojim matematickým intuíciám, podeliť sa o poznanie, a pomôcť si navzájom; dokážeme upraviť svoje prostredie, aby to pre nás bolo ľahšie, a vytvoriť nástroje, ktoré nám uľahčia veľa práce.

Naše skreslenia sú časťou nás samotných. Ale takisto je v nás aj tieň bayesovstva, omylné zariadenie, ktoré nás naozaj dokáže priviesť bližšie k pravde. Žiaden škriatok – ale predsa, nejaká pravda. Azda dosť na začiatok.

71 Luke Muehlhauser, „The Power of Agency,“ *Less Wrong (blog)* (2011), http://lesswrong.com/lw/5i8/-the_power_of_agency/.

E: Príliš pohodlné výhovorky

46. Správne použitie pokory

Všeobecne sa uznáva, že dobrá veda si vyžaduje nejaký druh pokory. *Aký druh*, to už je kontroverzné.

Predstavte si kreacionistu, ktorý hovorí: „Ale kto môže naozaj vedieť, či evolúcia je správna? Je to len teória. Mali by ste byť pokornejší a mať otvorenú myseľ.“ Je toto pokora? Evolucionista tu praktizuje veľmi prieberčivú nedostatočnú dôveru, odmieta zapojiť mohutné množstvá indícií v prospech záveru, ktorý mu je nepohodlný. Povedal by som, že či už toto nazvete slovom „pokora“ alebo nie, je to nesprávny krok v tanci.

A čo inžinier, ktorý pokorne navrhuje havarijné poistky do mechanizmu, hoci si je čertovsky istý, že mechanizmus sa nepokazí? Toto mi pripadá ako dobrý druh pokory. Z historického hľadiska sme už počuli o inžinieroch, ktorí si boli čertovsky istí, že sa nový stroj nepokazí, a potom sa aj tak pokazil.

A čo študent, ktorý pokorne robí skúšku správnosti na svojej písomke z matematiky? Opäť, toto by som označil ako dobrú pokoru.

A čo študent, ktorý povie: „Bez ohľadu na to, koľko skúšok správnosti urobím, nikdy si nemôžem byť *istý*, že moje odpovede v písomke sú správne,“ a preto neurobí ani jednu skúšku správnosti. Aj keď táto voľba pochádza z podobnej emócie ako u predchádzajúceho študenta, je menej múdra.

Odporučíte študovať usilovnejšie a študent odpovie: „Nie, to by u mňa nefungovalo; ja nie som jeden z tých chytrých ľudí ako vy; chudáčik ako ja nemôže dúfať v nič lepšie.“ Toto je spoločenská skromnosť, nie pokora. Súvisí s regulovaním postavenia v tlupe, nie s vedeckým procesom. Ak niekoho žiadate, aby bol „pokornejší“, štandardne si vaše slová spojí so spoločenskou skromnosťou – čo je intuitívny, každodenný, dedične daný pojem. Vedecká pokora je nedávny a zriedkavo používaný objav, a nie je vrodene spoločenská. Vedecká pokora je niečo, čo by ste mohli robiť aj sám v skafandri, vzdialený celé svetelné roky od Zeme, a nikto by sa nepozeral. Dokonca aj keby ste mali absolútnu záruku, že vás už nikto nikdy nebude kritizovať bez ohľadu na to, čo o sebe poviete alebo si pomyslíte. Aj tak by ste urobili skúšku správnosti vo výpočte, ak ste múdry.

Ten študent povie: „Ale ja som videl ako druhí študenti robia skúšky správnosti a potom sa aj tak ukázalo, že to majú zle. A vôbec, čo ak vďaka problému indukcie tentokrát 2+2=5? Bez ohľadu na to, čo urobím, nebude mať istotu.“ Znie to veľmi hlboko a veľmi pokorne. Ale nie je náhoda, že tento študent chce rýchlo odovzdať písomku, ísť domov a hrať počítačové hry.

Koniec nejakej éry vo fyzike sa zvyčajne neohlasuje hromom a trúbkami; častejšie začína niečím, čo vyzerá ako malá, maličká chyba... Ale keďže fyzici majú takú arogantnú predstavu, že ich modely by mali fungovať *vždy*, nielen *väčšinou*, všímajú si tieto malé chyby. Zvyčajne po podrobnejšom preskúmaní chybička zmizne. Len občas sa chyba rozšíri až do bodu, keď vyhodí do vzduchu celú teóriu. Preto je napísané: „Ak nehľadáš dokonalosť, zastavíš sa ešte pred prvým krokom.“

Pomyslite však na tú spoločenskú trúfalosť, snažiť sa mať *vždy* pravdu! Mám vážne podozrenie, že keby Veda tvrdila, že evolučná teória platí väčšinou, ale nie vždy – alebo keby Veda pripustila, že v niektoré dni Zem možno *je* plochá, veď kto to môže vedieť naisto – potom by vedci mali lepšiu povesť v spoločnosti. Veda by bola vnímaná ako menej konfrontačná, pretože by sme sa nemuseli hádať s ľuďmi, ktorí tvrdia, že Zem je plochá – bolo by dosť miesta na kompromis. Keď sa veľa hádate, ľudia vás vnímajú ako konfrontačných. Ak opakovane odmietate kompromisy, je to ešte horšie. Berte to ako otázku postavenia v tlupe: vedci si iste zaslúžili nejaké osobitné postavenie výmenou za také spoločensky užitočné nástroje ako lieky a mobilné telefóny. Avšak toto spoločenské postavenie neospravedlňuje ich požiadavku, aby sa na verejných školách učili *iba* vedecké myšlienky o evolúcii. Kňazi majú predsa takisto vysoké spoločenské postavenie. Vedci si žiadajú viac, než si zaslúžia – vyhrali trošku postavenia, a odrazu si myslia, že sú náčelníkmi celej tlupy! Mali by byť viac pokorní, a občas robiť kompromisy.

Zdá sa, že mnohí ľudia majú pomerne nejasné predstavy o „pokore rozumu“. Je nebezpečné mať normu, ktorej rozumiete iba nejasne; váš myšlienkový obraz môže mať toľko stupňov voľnosti, že sa

dokáže prispôsobiť, aby ospravedlnil takmer hocijaký skutok. Tam, kde ľudia majú nejasné myšlienkové modely, ktoré možno použiť na zdôvodnenie hocičoho, zvyčajne dospejú k tomu, že uveria tým veciam, ktorým už na začiatku veriť chceli. To je také pohodlné, že ľudia sa zvyčajne neradi vzdávajú neurčitosti. Ale cieľom našej etiky je, aby nás riadila, nie aby bola nami riadená.

„Pokora“ je často nepochopená cnosť. To neznamená, že by sme mali zahodiť pojem pokory, ale že by sme ho mali používať opatrne. Možno pomôže, ak sa pozrieme na *činy* odporúčané nejakým „pokorným“ typom myslenia a opýtame sa: „Robí ma takéto konanie silnejším alebo slabším?“ Ak myslíte na problém indukcie aplikovaný na most, ktorý má zostať stáť, dôjsť k záveru, že nič nie je isté, bez ohľadu na to, koľko bezpečnostných opatrení urobíte, môže znieť rozumne. Ak však uvážite rozdiely v skutočnom svete medzi pridaním pár káblov navyše alebo pokrčením plecami, zdá sa jasné, čo z toho robí most silnejším.

Prevažná väčšina odvolávok na „pokoru rozumu“, ktoré som videl, boli zámienkami na pokrčenie plecami. Ten, kto si kúpi žreb lotérie a hovorí: „Ale nemôžeš *vedieť*, že nevyhrám.“ Ten, kto neverí v evolúciu a hovorí: „Ale nemôžeš mi *dokázať*, že je to pravda.“ Ten, kto odmieta čeliť napohľad zložitému problému a hovorí: „Asi je príliš ťažké na vyriešenie.“ Problémom je motivovaný skepticizmus, čiže sklon nepotvrdzovať – skúmať oveľa prísnejšie tie tvrdenia, ktorým nechceme veriť. Pokora, ako je najčastejšie dezinterpretovaná, je úplne všeobecnou výhovorkou niečomu neveriť; vzhľadom na to, že aj tak si nemôžeme byť *istí*. Vyvarujte sa úplne všeobecných výhovoriek!

Ďalším problémom pokory je, že sa príliš ľahko *vyznáva*. V knihe *Breaking the Spell: Religion as a Natural Phenomenon*→ [„Odčarovanie: náboženstvo ako prírodný jav“] Dennett ukazuje, že hoci mnohým náboženským tvrdeniam je veľmi ťažké uveriť, pre človeka je ľahké uveriť, že by v ne *mal* veriť. Dennett toto nazýva „viera vo vieru“. Čo by to znamenalo naozaj predpokladať, naozaj veriť, že tri sa rovná jednej? Omnoho ľahšie je veriť, že by ste *mali* nejako veriť, že tri sa rovná jednej, a reagovať príslušne v primeraných chvíľach v kostole. Dennett naznačuje, že mnoho z „náboženskej viery“ by sa malo študovať ako „náboženské vyznávanie“ – to, čo si ľudia myslia, že by tomu mali veriť, a o čom vedia, že by to mali hovoriť.

Je príliš ľahké na každý protiargument odpovedať: „Nuž samozrejme, môžem sa mýliť.“ A potom, keď už ste si splnili svoju povinnosť zohnúť kolená smerom ku Skromnosti, keď ste jej vzdali vyžadovanú úctu, môžete pokračovať svojím pôvodným smerom a nezmeniť ani chlp.

Vždy je tu pokušenie tvrdiť čo najviac vecí s čo najmenšou námahou. Pokušenie spojiť všetky prichádzajúce informácie spôsobom, ktorý nám umožní zmeniť naše názory a najmä naše *konanie* čo najmenej. John Kenneth Galbraith povedal: „Keď má človek na výber medzi zmenou názoru a dôkazom, že názor zmeniť netreba, prakticky každý sa pustí do dokazovania.“[72] A čím *nepohodlnejšie* je zmeniť názor, tým viac úsilia ľudia na tento dôkaz vynaložia.

Ale viete, ak už ste tak či tak odhodlaní *urobiť* to isté, nemá zmysel sa takto neuveriteľne snažiť o zdôvodnenie. Často som zažil, ako ľudia dostali nové informácie, napohľad ich prijali, a potom starostlivo vysvetlili, prečo aj tak urobia presne tú istú vec, ktorú si predtým naplánovali, akurát s odlišným zdôvodnením. Zmyslom rozmýšľania je *utvárať* naše plány; ak sa chcete tak či tak držať pôvodných plánov, načo sa vôbec unúvať toľkým zdôvodňovaním? Keď sa stretnete s novou informáciou, tá ťažká časť je *aktualizovať*, *reagovať*, a nenechať túto informáciu len tak zmiznúť v čiernej diere. Nesprávne pochopená pokora je úžasnou čiernou dierou – jediné, čo treba urobiť, je pripustiť, že sa možno mýlite. Preto je napísané: „Byť pokorný znamená prijať konkrétne opatrenia v očakávaní vlastných chýb. Priznať svoju omylnosť a potom s ňou nič neurobiť nie je pokorné; je to chvastanie sa svojou skromnosťou.“

*→

→ https://books.google.sk/books/about/Breaking_the_Spell.html?id=yWtwDDqR61QC

72 John Kenneth Galbraith, *Economics, Peace and Laughter* (Plume, 1981), 50.

→ http://lesswrong.com/lw/gq/the_proper_use_of_humility/

47. Tretia možnosť

> Viera v Deda Mráza dáva deťom pocit úžasu a nabáda ich, aby sa pekne správali v nádeji, že dostanú darčeky. Ak bude viera v Deda Mráza zničená pravdou, deti stratia svoj pocit úžasu a prestanú sa pekne správať. Preto, hoci je viera v Deda Mráza fakticky nepravdivá, je to vznešená lož, ktorej čistý zisk by sme mali zachovať z utilitariánskych dôvodov.

Toto je klasicky nazývané falošná dilema, klam vylúčenia stredu, alebo klam výberu balíka. Aj keby sme prijali faktické a morálne predpokladu horeuvedeného argumentu, stále to nestačí. Ani predpoklad, že taktika Deda Mráza (povzbudzovanie detí, aby verili v Deda Mráza) je lepšia než nulová taktika (nerobiť nič), nedokazuje, že dedomrázizmus je *najlepšia zo všetkých možných alternatív*. Aj iné taktiky by mohli poskytnúť deťom pocit úžasu, napríklad vziať ich na sledovanie štartu Space Shuttle alebo zásobiť ich románmi science fiction. Podobne (ak si dobre spomínam), ponúkať deťom úplatky za dobré správanie povzbudzuje deti, aby sa správali dobre *iba* vtedy, keď sa dospelí pozerajú; zatiaľ čo chválenie bez úplatkov vedie k bezpodmienečne dobrému správaniu.

Vznešené lži bývajú vo všeobecnosti klamy výberu balíka; a odpoveďou na klam výberu balíka je, že ak jeho predpokladaný prínos naozaj potrebujeme, vieme si zostrojiť tretiu možnosť, ako ho získať.

Ako získame tretie možnosti? Prvým krokom k získaniu tretej možnosti je rozhodnúť sa, že ju ideme hľadať, a posledným krokom je rozhodnutie prijať ju. Znie to samozrejme, a predsa sa väčšina ľudí zasekne na týchto krokoch, a nie počas procesu hľadania. Odkiaľ pochádzajú falošné dilemy? Niektoré vznikajú poctivo, pretože je kognitívne náročné vidieť lepšie možnosti. Ale jednou továrňou na falošné dilemy je zdôvodňovanie pochybných rozhodnutí poukazovaním na predpokladané výhody oproti ničnerobeniu. V takomto prípade zdôvodňujúci *nechce* žiadnu tretiu možnosť; nájdenie tretej možnosti by zničilo toto zdôvodnenie. Posledná vec, ktorú chce dedomrázista počuť, je, že chvála funguje lepšie než úplatky, alebo že kozmická loď dokáže inšpirovať rovnako ako lietajúci sob.

Najlepšie býva nepriateľom dobrého. Ak je vaším cieľom *naozaj* pomáhať ľuďom, potom je lepšia možnosť dôvodom na oslavu – až nájdeme túto lepšiu stratégiu, budeme môcť ľuďom pomáhať účinnejšie. Ale ak je cieľom zdôvodniť jednu konkrétnu stratégiu *tvrdením, že pomáha ľuďom*, tretia možnosť je nepriateľský argument, konkurent.

Moderná kognitívna psychológia vníma rozhodovanie ako hľadanie možností. V skutočnom živote nestačí možnosti porovnávať, v prvom rade ich potrebujete vytvárať. Pri mnohých problémoch je možností tak veľa, že potrebujete kritérium, kedy sa v hľadaní zastaviť. Ak si chcete kúpiť dom, nemôžete porovnať každý dom v meste; v nejakom bude sa musíte prestať obzerať a rozhodnúť sa.

Ale čo keď sa naše vedomé motívy hľadania – kritériá, ktoré si môžeme pripustiť – nezhodujú s nevedomými vplyvmi? Keď vykonávame údajne altruistické hľadanie, hľadanie lepšej altruistickej taktiky, a nájdeme stratégiu, ktorá druhým pomáha, ale nás znevýhodňuje – nuž, neprestaneme hľadať *tam*; pokračujeme v hľadaní. Sami sebe však povieme, že hľadáme stratégiu, ktoré prinesie ešte väčší altruistický úžitok, samozrejme. Predpokladajme však, že nájdeme taktiku, ktorá má nejaký obhájiteľný úžitok, ale zhodou okolností *zároveň* vyhovuje aj nám osobne. Potom hľadanie ihneď ukončíme! V skutočnosti budeme pravdepodobne *odporovať* každej rade, aby sme opäť začali hľadať – vyhovoríme sa trebárs na nedostatok času. (Z nejakého dôvodu máme vždy kognitívne zdroje na hľadanie zdôvodnení našich existujúcich rozhodnutí.)

Dajte si pozor, keď si všimnete, že argumentuje v prospech rozhodnutia, ktoré je *obhájiteľné* namiesto *optimálneho*; alebo že má nejakú výhodu oproti ničnerobeniu, namiesto aby bolo najvýhodnejšou zo všetkých možných akcií.

Falošné dilemy sa časti vyskytujú ako zdôvodnenie neetických pravidiel, ktoré sú, obrovskou zhodou okolností, veľmi pohodlné. Klamanie je napríklad často omnoho pohodlnejšie než hovorenie pravdy; názor, ktorý ste mali na začiatku, je pohodlnejší než aktualizácia. Odtiaľ pochádza obľuba argumentov v prospech vznešených lží; slúžia ako obrana pôvodného názoru – nenájdete vznešeného

klamára, ktorý by vypočítal novú vznešenú lož; vždy si držia tú lož, s ktorou začali. Radšej toto hľadanie rýchlo ukončiť!

Ak to chcete robiť lepšie, opýtajte sa rovno: *Keby som videl, že existuje lepšia možnosť než moje terajšie rozhodnutie, tešil by som sa z hĺbky srdca, alebo by som cítil malý záblesk váhania, než by som sa vzdal?* Ak sú odpovede „nie“ a „áno“, dajte si pozor na to, že ste asi nehľadali tretiu možnosť.

Čo vedie k ďalšej dobre otázke, ktorú môžete sami sebe položiť: *Strávil som päť minút so zavretými očami brainstormujúc divoké a tvorivé možnosti, snažiac sa vymyslieť lepšiu možnosť?* Musí to byť päť minút skutočného času, lebo inak by ste iba žmurkli – zavreli oči a opäť ich otvorili – a povedali: „Veru hej, hľadal som iné možnosti, ale žiadne nie sú.“ Žmurknutie je dobrou čiernou dierou, do ktorej môžete odhodiť svoje povinnosti. Odporúčam použiť skutočné, fyzické hodiny.

A čo sa týka tých divokých a tvorivých možností – dávali ste si dobrý pozor na to, aby ste nevymysleli niečo dobré? Mali ste v kútiku mysle tajné úsilie zaručiť, aby každá možnosť, nad ktorou sa zamyslíte, bola očividne zlá?

Je úžasné, ako veľa vznešených klamárov a im podobných je ochotných prijať etické priestupky – pri všetkom povinnom plači nad svojou agóniou svedomia – hoci nestrávili ani päť minút na hodinách hľadaním inej možnosti. Existujú myšlienkové hľadania, pri ktorých si tajne želáme, aby sme neuspeli; a keď je vidina úspechu nepohodlná, ľudia prijmú prvú dostupnú výhovorku, aby sa vzdali.

* →

48. Lotérie: Plytvanie nádejou

Klasická kritika lotérií znie, že ľudia, ktorí si kupujú žreby, sú práve tí, ktorí si najmenej môžu dovoliť prehrať; že lotéria je bezodnou jamou na peniaze, v ktorej sa stráca majetok práve tých, ktorí ho najviac potrebujú. Niektorí zástancovia lotérie, dokonca aj niektorí diskutéri na *Overcoming Bias*, sa pokúšali obhajovať kupovanie žrebov lotérie ako *rozumný nákup peknej predstavy* – zaplatíte jeden dolár za jeden deň príjemného očakávania, keď si sami seba predstavujete ako milionára.

Ale zamyslite sa, čo presne z toho vyplýva. Znamenalo by to, že zapĺňate svoj vzácny mozog predstavou, ktorej reálna pravdepodobnosť je takmer nulová – taký úzky slížik pravdepodobnosti, že si ho ani len nedokážete predstaviť. Loptičky v osudí rozhodnú o vašom osude. Je to predstava bohatstva, ktoré príde bez úsilia – bez usilovnosti, učenia, šarmu, dokonca aj bez trpezlivosti.

To robí z lotérie iný druh plytvania: plytvanie emocionálnou energiou. Podporuje ľudí, aby investovali svoje sny, svoje nádeje na lepšiu budúcnosť, do nekonečne malej pravdepodobnosti. Nebyť lotérie, možno by si predstavovali, že pôjdu na technickú školu, že si založia vlastnú firmu, že dosiahnu povýšenie v práci – veci, ktoré by mohli dokázať naozaj *urobiť*; nádeje, ktoré by mohli spôsobiť, že budú chcieť byť silnejší. Ich snívajúci mozog by si mohol pri 20-tom vizualizovaní príjemnej predstavy všimnúť spôsob, ak to naozaj urobiť. Vari nemáme mozgy práve *na toto*? Ale ako môže takýto počin v hraniciach skutočnosti súperiť s umelo sladenou vyhliadkou na okamžité bohatstvo – nie po dotiahnutí dotcomového startupu na burzu, ale hneď tento utorok?

Vážne, prečo nemôžeme skrátka povedať, že kupovať žreby lotérie je hlúpe? Ľudia *sú* z času na čas hlúpi – nemala by to byť taká prekvapujúca hypotéza.

Nie je prekvapením, že ľudský mozog nerobí 64-bitovú aritmetiku, nedokáže emocionálnu silu príjemného očakávania vynásobiť činiteľom 0,000 000 01 a nestratiť sa pritom. Nie je prekvapením, že mnohí ľudia si neuvedomujú, že matematický výpočet očakávaného úžitku by mal *nahradiť* ich nepresné finančné inštinkty, a namiesto toho berú tento výpočet ako púhy ďalší *argument*, ktorý treba postaviť proti ich príjemnému očakávaniu – ako emocionálne slabý argument, keďže je výsledkom púheho čmárania na papier, namiesto vízie rozprávkového bohatstva.

Toto mi pripadá dostačujúce na vysvetlenie obľúbenosti lotérií. Prečo toľko diskutérov cíti nutkanie brániť túto klasickú formu samozničenia?

→ http://lesswrong.com/lw/hu/the_third_alternative/

Proces prekonávania skreslení si vyžaduje (1) všimnúť si dané skreslenie, (2) podrobne ho analyzovať, (3) rozhodnúť sa, že je zlé, (4) zistiť, ako sa mu vyhnúť, a potom (5) to naozaj urobiť. Je smutné, koľko ľudí sa dostane cez kroky 1 a 2 a potom sa zaseknú na kroku 3, ktorý by právom mal byť zo všetkých piatich ten najjednoduchší. Nemali by sme sa nasilu pokúšať hľadať na skresleniach niečo dobré – proste sa ich *zbavme*.

⁂→

49. Nová vylepšená lotéria

Ľudia stále naznačujú, že lotéria nie je plytvanie nádejou, ale služba, ktorá umožňuje nákup fantázie - „denné snívanie o staní sa milionárom, omnoho lacnejšie než denné snívanie o hollywoodskych filmových hviezdach“. Jeden diskutér napísal: „Je veľký rozdiel medzi nulovou šancou zbohatnúť, a epsilonovou. Kúpa žrebu umožňuje tvojmu snu o bohatstve preklenúť túto medzeru.“

V skutočnosti, jednou z poínt, ktoré som sa pokúšal urobiť, je že medzi nulovou šancou na zbohatnutie a epsilonovou šancou, rozdiel je rádovo epsilon. Ak o tom pochybujete, nech sa epsilon rovná $1/10^{10^{100}}$.

A vôbec: Ak chceme predstierať, že lotéria predáva epsilonovú nádej, mali by sme vymyslieť Novú Vylepšenú Lotériu. Nová Vylepšená Lotéria vypláca výhry priemerne raz za päť rokov, v náhodnom čase – určenom povedzme rozpadom nie veľmi rádioaktívneho prvku. Jedenkrát si kúpite žreb, za jediný dolár, a dostanete nielen pár dní epsilonovej šance zbohatnúť, ale niekoľko *rokov* epsilonu. A nielen to; vaše bohatstvo môže prísť v ľubovoľnom okamihu! V *hociktorú minútu* môže zazvoniť telefón a oznámiť vám, že *vy, áno, vy* ste sa stali milionárom!

Pomyslite, o čo lepšie by bolo toto oproti bežnému žrebovaniu lotérie, ktoré sa robí iba v určenú dobu, niekoľkokrát za týždeň. Povedzme, že príde šéf a chce, aby ste prerobili návrh, priniesli zásoby zo skladu, alebo niečo rovnako otravné. Namiesto pustenia sa do práce by ste sa mohli otočiť a zahľadieť na telefón, dúfajúc že príde ten hovor – pretože by existovala epsilonová šanca, že *práve v tej chvíli práve vy, áno, vy* získate hlavnú cenu! A keby sa to aj nestalo *túto* minútu, nie je dôvod byť sklamaný – môže sa to stať *nasledujúcu* minútu!

Pomyslite, koľko veľa nových fantázií by táto Nová Vylepšená Lotéria umožnila. Mohli by ste nakupovať v obchode, ukladať drahé tovary do svojho nákupného vozíka – ak váš mobil nezazvoní so správou o vyhranej lotérii, stále ešte môžete tieto tovary vyložiť, nie?

Možno by Nová Vylepšená Lotéria mohla ukazovať aj ustavične kolísajúcu pravdepodobnostnú distribúciu možnosti výhry, a možnosti vyžrebovanie konkrétnych čísel, s celkovým očakávaním súhrnne sa rovnajúcim spomínanému Poissonovmu rozdeleniu. Pomyslite, aké zábavné by bolo *toto*! Ty brďo, práve túto minútu je šanca vyhrať desaťkrát vyššia než zvyčajne! A pozri, číslo 42, ktoré som si vybral ako šťastné číslo, má takmer dvojnásobok bežnej šance vyhrať! Mohli by ste to občanom vysielať na displeje mobilov, aby stačilo otvoriť telefón a uvidieť šancu na víťazstvo. Pomyslite, aké vzrušujúce by bolo *toto*! Omnoho vzrušujúcejšie než plánovanie s vkladnou knižkou! Omnoho vzrušujúcejšie než písanie domácej úlohy! Tento nový sen by bol o toľko sladší, že by nekonkuroval iba nádeji dostať sa na vysokú školu technického smeru, ale dokonca aj nádeji na skorý návrat domov z práce. Ľudia by mohli jednoducho zostať prilepení k obrazovke po celý deň, nepotrebovali by predsa snívať o ničom *inom*!

Veď hej, poskytovať ľuďom lákavé sny, *ktoré sa v skutočnosti nesplnia*, je iste hodnotná služba, máte pravdu. Ľudia sú za ňu ochotní platiť, no tak musí byť hodnotná. Alternatíva by znela, že zákazníci sa mýlia, a ako všetci vieme, to je nemožné.

A napriek tomu dnešné vlády so svojím hnusným monopolom na lotérie neponúkajú túto jednoduchú a očividnú službu. Prečo? Pretože chcú z ľudí vyžmýkať čo najviac. Chcú, aby míňali peniaze každý týždeň. Chcú, aby míňali stovky dolárov za vzrušenie z viery, že ich šanca vyhrať je stokrát väčšia, namiesto možnosti pozerať na obrazovku mobilu a čakať, kedy pravdepodobnosť narastie.

→ http://lesswrong.com/lw/hl/lotteries_a_waste_of_hope/

Ak teda veríte, že lotéria je služba, je to očividne nenormálne predražená služba – za ktorú platia najchudobnejší členovia našej spoločnosti – a vašou posvätnou občianskou povinnosťou žiadať si namiesto toho Novú Vylepšenú Lotériu.

*→

50. Ale stále je tu šanca, nie?

Pred rokmi som hovoril s niekým, kto mimochodom poznamenal, že neverí v evolúciu. Ja som povedal: „Toto nie je devätnáste storočie. Keď Darwin prvýkrát navrhol evolúciu, mohlo byť rozumné pochybovať o nej. Ale toto je dvadsiate prvé storočie. Dokážeme *čítať gény*. Ľudia a šimpanzy majú spoločných 98 % DNA. Vieme, že ľudia a šimpanzy sú príbuzní. Je to *jasné*."

On: „Možno je tá DNA podobná iba náhodou?"

Ja: „Pravdepodobnosť niečoho takého je asi dve na sedemsto päťdesiat miliónu k jednej."

On: „Ale stále je tu šanca, nie?"

Dobre, je tu pár dôvodov, prečo si moje minulé ja nemôže v tejto konverzácii nárokovať jasné morálne víťazstvo. Jeden z dôvodov je, že si nepamätám, odkiaľ som vzal ten vzorec $2^{750000000}$, ale pravdepodobne je to meta-rádovo sedí. Druhým dôvodom je, že moje minulé ja nepoužilo pojem kalibrovanej istoty. Zo všetkých prípadov v histórii ľudstva, keď človek vypočítal, že niečo má šancu $2^{750000000}:1$, nepochybne sa mýlil vo viac než jednom z $2^{750000000}$ prípadov. Napríklad ten odhad spoločných génov bol opravený na 95 %, nie 98 % - a aj to sa možno týka iba 30 000 známych génov a nie celého genómu, a v tom prípade to ani meta-rádovo nesedí.

Ale myslím si, že odpoveď toho druhého je stále veľmi smiešna.

Nespomínam si, čo som povedal potom – asi niečo ako „**Nie**" - ale pamätám si tento prípad, pretože mi dal niekoľko vhľadov do zákonov myslenia, ako im rozumejú neosvietení.

Prvýkrát mi napadlo, že ľudská intuícia robí kvalitatívny rozdiel medzi „Nie je šanca" a „Veľmi maličká šanca, na ktorú sa oplatí myslieť". Vidíte v debate o lotérii na *Overcoming Bias*, kde niekto povedal: „Je veľký rozdiel medzi nulovou šancou na výhru a epsilonovou šancou na výhru," a ja som odpovedal: „Nie, ten rozdiel je v ráde epsilonov; ak o tom pochybuješ, nech sa epsilon rovná $1/10^{10^{100}}$."

Problém je, že teória pravdepodobnosti nám občas dovolí vypočítať šancu, ktorá je naozaj príliš malá na to, aby bola hodna myšlienkového priestoru, ktorý zaberá – ale v tom čase už ste ju vypočítali. Ľudia si mýlia mapu a územie, takže na inštinktívnej úrovni myslenie na symbolicky opísanú pravdepodobnosť sa zdá ako „pravdepodobnosť, na ktorú sa oplatí myslieť", aj keď tento symbolický popis *odkazuje na* číslo také malé, že keby to bolo zrnko prachu, nevideli by ste ho. Dokážeme použiť slová na opísanie takýchto malých čísel, ale nie pocity – takýto malý pocit neexistuje, neaktivoval by dostatok neurónov a neuvoľnil by dostatok neurotransmiterov, aby sme ho cítili. To je dôvod, prečo ľudia kupujú žreby v lotérii – nikto nedokáže *precítiť* malosť takejto malej pravdepodobnosti.

Ale čo mi pripadá ešte fascinujúcejšie, je kvalitatívne rozlišovanie medzi „istými" a „neistými" argumentmi, kde pokiaľ argument nie je istý, máte povolené ho ignorovať. Napríklad, ak je pravdepodobnosť nula, musíte sa daného názoru vzdať, ale ak je pravdepodobnosť $1/10^{100}$, môžete si ho ponechať.

Dobre, toto je slobodná krajina a nikto by vás nemal posadiť za mreže kvôli nelegálnemu rozmýšľaniu, ale ak ste sa rozhodli ignorovať argument, ktorý hovorí, že pravdepodobnosť je $1/10^{100}$, prečo neignorovať aj argument, ktorý hovorí, že pravdepodobnosť je nula? Chcem tým povedať, keďže ste sa tak či tak rozhodli ignorovať indície, prečo je o toľko horšie ignorovať isté indície než neisté indície?

Často som v živote zistil, že som sa poučil z do očí bijúcich zlých príkladov druhých ľudí, keď som ich zovšeobecnil na menej jasné prípady. V tomto prípade je doplňujúcim ponaučením to, že ak nemôžete

→ http://lesswrong.com/lw/hm/new_improved_lottery/

ignorovať pravdepodobnosť $1/10^{100}$ len preto, že sa vám zachcelo, nemôžete ignorovať ani pravdepodobnosť 0,9 preto, že sa vám zachcelo. Je to ten istý šmykľavý svah.

Spomeňte si na tento príklad, ak sa niekedy pristihnete, že si myslíte: „Ale nemôžete mi *dokázať*, že sa mýlim.“ Ak ste rozhodnutí ignorovať pravdepodobnostný argument, prečo neignorovať aj dôkaz?

*→

—

51. Klam sivej

> Sofistik: „Svet nie je čiernobiely. Nikto nekoná čisté dobro ani čisté zlo. Všetko je sivé. Preto nikto nie je lepší než nikto iný.“
>
> Zetét: „Keďže poznáš iba sivú, došiel si k záveru, že všetky sivé majú rovnaký odtieň. Vysmievaš sa z jednoduchosti dvojfarebného pohľadu, ale sám si ho nahradil jednofarebným...“
>
> --Marc Stiegler, Dávidov prak[73]

Neviem, či táto Sofistikova chyba má oficiálny názov, ale ja ju nazývam klam sivej. Videli sme ju vo včerajšom článku – ten, ktorý veril, že šanca dve na sedemsto päťdesiat miliónu k jednej proti nemu znamená „stále je tu šanca“. Všetky pravdepodobnosti boli preňho jednoducho „nejasné“ a to znamenalo, že má právo ignorovať ich, ako sa mu zachce.

„Mesiac sa skladá zo zeleného syra“ a „Slnko sa skladá najmä z vodíka a hélia“ sú oba neisté výroky, ale nemajú rovnakú neistotu.

Všetko sú odtiene sivej, ale existujú také svetlé odtiene sivej, že je to takmer biela, a také tmavé odtiene sivej, že je to takmer čierna. A aj keby nie, stále môžeme porovnávať odtiene a hovoriť „tento je tmavší“ a „tento je svetlejší“.

Pred rokmi bolo jedným z čudných formujúcich momentov mojej kariéry racionalistu prečítanie tohto odseku z knihy *Hráč hier* od Iaina M. Banksa, najmä zvýraznená veta:[74]

> Systém viny nerozoznáva žiadnych nevinných. Ako pri každom mocenskom aparáte, ktorý si myslí, že každý je buď zaňho alebo proti nemu, my sme proti nemu. Aj ty by si bol, keby si nad tým rozmýšľal. Samotný spôsob, akým rozmýšľaš, ťa radí medzi jeho nepriateľov. To nemusí byť tvoja chyba, pretože **každá spoločnosť vtláča niektoré svoje hodnoty tým, ktorí v nej vyrastajú, ale pointa je, že niektoré spoločnosti sa tento účinok snažia maximalizovať a niektoré minimalizovať.** Ty pochádzaš z tej druhej a žiadajú ťa, aby si sa vysvetlil niekomu z tej prvej. Vykrúcanie sa bude omnoho ťažšie než si dokážeš predstaviť; neutralita je pravdepodobne nemožná. Nemôžeš si vybrať nemať postoje, ktoré máš; nie sú to nejaké oddelené veci, ktoré by si nejako mohol odpojiť od zvyšku tvojej bytosti; sú funkciou tvojej existencie. Ja to viem a oni to vedia; mal by si sa s tým radšej zmieriť.

Teraz mi nepíšte zlostné komentáre, že keď spoločnosti vtláčajú menej svojich hodnôt, potom každá nasledujúca generácia musí viac pracovať, aby sa dostala zo štartu. To nie je to, čo som si z tohto odseku odniesol.

Z tohto odseku som si odniesol niečo, čo v retrospektíve vyzerá natoľko zrejmé, že som to mohol nájsť na stovkách miest; ale niečo v tomto odseku spôsobilo, že mi to docvaklo.

Bol to celý pojem Kvantitatívnej Cesty aplikovanej na problémy života ako sú morálne úsudky a hľadanie osobného sebazdokonaľovania. Že aj keď nedokážete niečo len tak zapnúť a vypnúť, aj tak budete mať sklon to zvyšovať alebo znižovať.

→ http://lesswrong.com/lw/ml/but_theres_still_a_chance_right/

73 Marc Stiegler, *David's Sling* (Baen, 1988).

74 Iain Banks, *The Player of Games* (Orbit, 1989).

Je to také samozrejme, že sa to neoplatí spomínať? Hovorím, že to nie je príliš samozrejmé, pretože nejeden bloger povedal o *Overcoming Bias*: „Je to nemožné, nikto nemôže úplne odstrániť skreslenia.“ Nezaujíma ma, či je dotyčný profesionálny ekonóm, je jasné, že ešte nepochopil, ako sa Kvantitatívna Cesta týka každodenného života a vecí ako osobné sebazdokonaľovanie. To, čo nedokážem *odstrániť*, sa mi ešte stále môže oplatiť *redukovať*.

Alebo si vezmite túto výmenu medzi Robinom Hansonom a Tylerom Cowenom. Robin Hanson povedal, že sa snaží dať aspoň 75 % váhy predpisom ekonomickej teórii oproti vlastnej intuícii: „Snažím sa väčšinou iba priamočiaro aplikovať ekonomickú teóriu, a pridávať k nej iba málo osobného alebo kultúrneho hodnotenia.“ Tyler Cowen odpovedal:

> Podľa mňa neexistuje nič také ako „priamočiaro aplikovať ekonomickú teóriu“... teórie sa vždy aplikujú cez naše osobné a kultúrne filtre a nijako inak to nemôže byť.

Áno, ale môžete sa snažiť minimalizovať tento efekt, alebo môžete robiť veci, ktoré ho iste zvýšia. A *ak* sa ho snažíte minimalizovať, potom si myslím, že v mnohých prípadoch nie je nerozumné nazvať výstup „priamočiarym“ - dokonca aj v ekonómii.

„Každý je nedokonalý.“ Mohandas Gándhí bol nedokonalý, aj Josif Stalin bol nedokonalý, ale nebol to rovnaký odtieň nedokonalosti. „Každý je nedokonalý!“ je výborný príklad nahradenia dvojfarebného pohľadu jednofarebným. Ak poviete: „Nikto nie je dokonalý, ale *niektorí ľudia sú menej nedokonalí ako iní*,“ možno nezožnete potlesk; ale tým, ktorí sa snažia robiť veci lepšie, ste ukázali nádej. Napokon, nikto nie je *dokonale* nedokonalý.

(Vždy keď mi niekto povie: „Perfekcionizmus ti škodí,“ odpoviem: „Myslím si, že je okej byť nedokonalý, ale nie taký nedokonalý, že si to druhí ľudia všimnú.“)

Podobná je hlúposť tých, ktorí hovoria: „Každá vedecká paradigma prináša nejaké svoje predpoklady, podľa ktorých interpretuje pokusy,“ a potom sa správajú, ako by tým dokázali, že veda je na rovnakej úrovni ako šamanstvo. Každý svetonázor vtláča niečo zo svojej štruktúry do svojich pozorovaní, ale pointa je, že existujú svetonázory, ktoré sa snažia toto vtláčanie minimalizovať, a sú svetonázory, ktoré sa v tom vyžívajú. Neexistuje biela, ale existujú odtiene sivej, ktoré sú omnoho svetlejšie než iné, a je hlúposť zaobchádzať s nimi, akoby boli všetky na rovnakej úrovni.

Ak mesiac posledných pár miliárd rokov obiehal okolo Zeme, ak ste ho posledných pár rokov videli na oblohe, a ak očakávate, že ho uvidíte na jeho určenom mieste v jeho určenej fáze aj zajtra, toto nie je istota. A ak očakávate, že neviditeľný drak vylieči vašej dcére rakovinu, ani toto nie je istota. Ale sú to pomerne odlišné stupne neistoty – očakávanie, že sa veci opäť stanú rovnakým spôsobom, aký ste v minulosti predpovedali na dvanásť desatinných miest, verzus očakávanie, že sa stane niečo, čo *porušuje* doteraz pozorovaný poriadok. Nazvať obidvoje „vierou“ sa mi zdá trochu príliš nepresné.

Je to veľmi zaujímavé, psychologicky – celá tá vec, že: „Aj veda je založená na viere, tak vidíš!“ Typicky to hovoria ľudia, ktorí tvrdia, že viera je *dobrá* vec. Prečo teda hovoria: „Aj veda je založená na viere!“ tým zlostným triumfálnym tónom, a nie ako kompliment? A to pomerne *nebezpečný* kompliment, človek by si pomyslel, z ich pohľadu. Ak je veda založená na „viere“, potom je veda rovnakého typu ako náboženstvo – možno ich priamo porovnávať. Ak je veda náboženstvom, potom je to náboženstvo, ktoré uzdravuje chorých a odhaľuje tajomstvá hviezd. Dávalo by zmysel povedať: „Kňazi vedy dokážu jasne, verejne, overiteľne chodiť po Mesiaci ako zázrak založený na viere, a kňazi tvojej viery toto nedokážu.“ Ste si istí, že chcete ísť týmto smerom, veriaci? Keď sa nad tým lepšie zamyslíte, možno by ste celé to „Aj veda je náboženstvo!“ radšej odvolali.

Je to zvláštna dynamika: Pokúšate sa vyčistiť svoj odtieň sivej a dostanete ho do bodu, kde je pomerne svetlý, načo niekto vstane a povie hlboko urazeným tónom: „Ale to nie je biela! Je to sivá!“ Jedna vec je, keď niekto povie: „Nie je to také svetlé ako si myslíš, pretože sú tu konkrétne problémy X, Y a Z.“ Ale iná vec je, keď niekto zlostne povie: „To nie je biela! Je to sivá!“ bez poukázania na konkrétne tmavé škvrny.

V tomto prípade začnem podozrievať, že sa jedná o psychológiu ešte nedokonalejšiu než je bežné – že niekto možno uzavrel diabolskú zmluvu so svojimi vlastnými chybami a teraz odmieta počuť

o hocijakej možnosti zlepšenia. Keď si niekto nájde výhovorku, prečo sa nepokúšať zlepšiť, často odmieta pripustiť, že by sa niekto iný *mohol* pokúsiť zlepšiť, a každý spôsob zlepšenia je odvtedy jeho nepriateľom, a každé tvrdenie, že je možné pohnúť sa dopredu, je preňho urážkou. A tak jedným dychom pyšne povedia: „Som rád, že som sivý," a nasledujúcim dychom zlostne: „A *ty si tiež sivý!*"

Ak neexistuje čierna a biela, stále existuje svetlejšia a tmavšia, a nie všetky sivé sú rovnaké.

G2 pripomína Asimovovu Relativitu nesprávneho:[75] „Keď si ľudia mysleli, že zem je plochá, mýlili sa. Keď si ľudia mysleli, že zem je guľa, mýlili sa. Ale ak si myslíš, že myslieť si, že zem je guľatá je rovnako nesprávne ako myslieť si, že zem je plochá, potom je tvoj pohľad ešte nesprávnejší než oni obaja dohromady."

*→

52. Absolútna autorita

Príde k vám niekto a povýšenie povie: „Veda vlastne nič naozaj *nevie*. Máte akurát *teórie* – nemôžete vedieť *naisto*, že máte pravdu. Vy vedci ste zmenili názor na to, ako funguje gravitácia – čo ak zajtra rovnako zmeníte názor na evolúciu?"

Hľaďte na tú hlbočiznú kultúrnu priepasť. Ak si myslíte, že ju prekleniete niekoľkými vetami, čaká vás trpké sklamanie.

Vo svete neosvietených existujú iba autority a neautority. Čomu možno veriť, tomu možno veriť; čomu nemožno veriť, to môžete v pohode zahodiť. Existujú dobré zdroje informácií a zlé zdroje informácií. Ak vedci čo len raz vo svojej histórii zmenili svoje príbehy, potom veda nemôže byť skutočnou Autoritou a už nikdy jej nemožno dôverovať – ako keď pristihnete svedka pri rozpore vo výpovedi, alebo keď prichytíte zamestnanca ako kradne z pokladne.

Navyše, dotyčný považuje za samozrejmé, že zástanca nejakej myšlienky ju musí obhajovať proti všetkým možným protiargumentom a nesmie nič pripustiť. Všetky tvrdenia sú hodnotené podľa toho. Ak ešte aj *zástanca* vedy pripúšťa, že veda je menej než dokonalá, potom musí byť úplne bezcenná.

Keď niekto prežil celý svoj život zvyknutý na istotu, nemôžete mu len tak povedať: „Veda je pravdepodobnostná, rovnako ako všetko iné poznanie." Prvú polovicu vašej vety vezme ako priznanie viny; druhú polovicu odmietne ako chabý pokus obviniť všetkých ostatných, aby ste sa vyhli odsúdeniu.

Priznali ste, že nie ste dôveryhodní – choď teda preč, veda, a viac nás neobťažuj!

Jedným zrejmým zdrojom tohto myšlienkového vzorca je náboženstvo, ktorého písma údajne pochádzajú od Boha; preto by priznanie ľubovoľnej chyby v nich úplne zničilo ich autoritu; preto je každá stopa pochybnosti hriechom, a vyhlasovať istotu je *povinné* bez ohľadu na to, či ste si istí alebo nie.

Ale obávam sa, že tradičný školský systém je tu tiež na vine. Učiteľ vám povie isté veci a vy im musíte veriť a musíte ich na skúške naspäť odrecitovať. Keď však niečo v triede povie študent, nemusíte sa tomu podriadiť – (zdá sa, že) máte slobodu súhlasiť alebo nesúhlasiť a nikto vás za to nepotrestá.

Obávam sa, že táto skúsenosť mapuje oblasť viery na spoločenské oblasti *autority, príkazov, zákona*. V spoločenskej oblasti je kvalitatívny rozdiel medzi absolútnymi zákonmi a neabsolútnymi zákonmi, medzi príkazmi a odporúčaniami, medzi autoritami a neautoritami. Zdá sa, že existuje presné poznanie a nepresné poznanie, tak ako presné predpisy a nepresné predpisy. Presným autoritám sa treba podrobiť, kým nepresné odporúčania možno poslúchať alebo ignorovať podľa osobného rozhodnutia. A veda, keďže sama pripúšťa, že má možnosť chyby, musí patriť do tej druhej kategórie.

(Na okraj poznamenávam, že vidím určitú podobnosť s tými, ktorí si myslia, že ak nedostanete autoritatívnu pravdepodobnosť napísanú na kúsku papiera od učiteľa v triede, alebo z nejakého iného nespochybniteľného zdroja, potom vaša neistota nie je vecou bayesovskej teórie pravdepodobnosti. Niekto by mohol – *ach!* - nesúhlasiť s vaším odhadom pôvodnej pravdepodobnosti. Tak sa zdá tým, ktorí

75 Isaac Asimov, *The Relativity of Wrong* (Oxford University Press, 1989).

→ http://lesswrong.com/lw/mm/the_fallacy_of_gray/

nie sú celkom osvietení, že bayesovské pôvodné pravdepodobnosti patria do kategórie názorov, ktoré navrhujú študenti, a nie do kategórie názorov, ktoré prikazujú učitelia – nie je to pravé *poznanie*.)

Táto hlbočizná kultúrna priepasť medzi autoritatívnou cestou a kvantitatívnou cestou je veľmi otravná pre tých, ktorí ponad ňu hľadia z racionalistickej strany. Je tu niekto, kto verí, že má poznanie *spoľahlivejšie* než je púhy pravdepodobnostný odhad vedy – ako sú odhady, že mesiac zajtra vyjde na svojom určenom mieste a v určenej fáze, tak ako to bolo každú pozorovanú noc od vynálezu astronomických záznamov, a ako to predpovedajú fyzikálne teórie, ktorých predchádzajúce predpovede boli úspešne potvrdené na štrnásť desatinných miest. A čo je to poznanie, ktoré ten neosvietený vyzdvihuje nad to naše, a prečo? Pravdepodobne je to nejaký plesnivý starý zvitok, ktorý bol vyvrátený jedenástimi spôsobmi v nedeľu, v pondelok, a v každý ďalší deň v týždni. Napriek tomu je (podľa ich slov) spoľahlivejší než veda, pretože si nikdy nepripustí chybu, nikdy nezmení názor, bez ohľadu na to, koľkokrát ho vyvrátia. Pohadzujú slovom „istota“ ako tenisovou loptou, používajú ho zľahka ako pierko – zatiaľ čo vedci sú zaťažení povinnosťou pochybovať, a zápasia, aby dosiahli aspoň štipku pravdepodobnosti. „Ja som dokonalý,“ hovoria bez ohľadu na svet, „iste som veľmi vysoko nad vami, ktorí sa stále musíte namáhať, aby ste sa zdokonalili.“

Nie je nič jednoduché, čo by ste im mohli povedať – žiadna *rýchla* zdrvujúca odpoveď. Keď pozorne uvažujete, možno sa vám podarí získať si obecenstvo, ak je to verejná diskusia. Nanešťastie však nemôžete len tak vyhŕknuť: „Hlúpy smrteľník, kvantitatívna cesta je za hranicami tvojho chápania, a názory, ktoré tak ľahkovážne nazývaš ‚istoty‘ sú menej isté než tá najmenšia z našich mocných hypotéz.“ Je to rozdiel celkového *životného nastavenia*, ktorý nie je ľahké opísať slovami a už vôbec nie rýchlo.

Čo by ste mohli povedať, ako rečník pred obecenstvom? Ťažko povedať... možno:

- „Sila vedy pochádza zo schopnosti zmeniť názor a priznať si, keď sa mýlime. Ak si vy nikdy nepriznáte, že sa mýlite, to ešte neznamená, že robíte menej chýb.“
- „Každý môže *povedať*, že si je absolútne istý. Ťažšie je však nikdy, naozaj nikdy neurobiť žiadnu chybu. Vedci si tento rozdiel uvedomujú, a preto nehovoria, že sú si absolútne istí. To je celé. Neznamená to, že majú nejaký konkrétny dôvod pochybovať o nejakej teórii – aj keby každý kúsok indície ukazoval tým istým smerom, keby sa všetky hviezdy a planéty zoradili ako dominové kocky, aby podporili jednu hypotézu, vedci aj tak nepovedia, že sú si absolútne istí, pretože skrátka majú vyššie kritériá. To neznamená, že by vedci mali menší *nárok* na istotu než povedzme politici, ktorí sú si vždy takí istí všetkým.“
- „Vedci nepoužívajú slová ‚nie som si absolútne istý‘ takým spôsobom, ako ste zvyknutí v bežnej konverzácii. Chcem tým povedať, predstavte si, že by ste boli u lekára na krvnom teste, a lekár by prišiel a povedal: ‚Urobili sme nejaké testy, ale nie je absolútne isté, že sa neskladáte zo syra, a je nenulová šanca, že dvadsať víl zložených z inteligentnej čokolády spieva pesničku od Barneyho „I love you“ vo vašom hrubom čreve.‘ Utekajte odušu, váš lekár sám potrebuje lekára. Keď však vedec povie to isté, znamená to, že si myslí, že tá pravdepodobnosť je taká malá, že by ste ju neuvideli ani v elektrónovom mikroskope, ale je ochotný pozrieť sa na vaše indície v tom extrémne nepravdepodobnom prípade, že by ste nejaké mali.“
- „Boli by ste ochotní zmeniť svoj názor na niektoré veci, ktoré nazývate ‚isté‘, keby ste videli dosť indícií. Myslím tým, predstavte si, že by sa samotný Boh zniesol z oblakov a povedal by vám, že celé vaše náboženstvo je pravdivé, s výnimkou narodenia z panny. Ak by toto dokázalo zmeniť váš názor, potom nemôžete povedať, že ste si absolútne istí narodením z panny. Z technických dôvodov teórie pravdepodobnosti, ak je teoreticky možné, že by ste o niečom zmenili svoj názor, potom to nemôže mať pravdepodobnosť presne rovnú jednej. Táto neistota môže byť menšia než zrnko prachu, ale musí tam byť. A ak by ste nedokázali zmeniť názor ani keby vám Boh povedal, že je to inak, potom máte problém pripustiť, že sa mýlite, ktorý zrejme prekračuje všetko, čo vám smrteľník ako ja môže povedať.“

V istom zmysle je však zaujímavejšia otázka, čo povedať niekomu, keď *nie ste* pred obecenstvom. Ako začať ten dlhý proces učenia niekoho, ako žiť vo vesmíre bez istoty?

Myslím si, že prvým krokom by malo byť pochopenie, že bez istoty sa žiť *dá* – že aj *keby* ste si, *hypoteticky povedané*, neboli ničím istí, to vám stále neberie schopnosť robiť morálne alebo faktické rozdiely. Parafrázujúc Loisa Bujolda: „Netlač silnejšie, znižuj odpor."

Jedna z častých *obhajob* Absolútnej Autority je niečo, čo nazývam: „argumentovanie argumentovaním sivou", a vyzerá to asi takto:

- Morálny relativisti hovoria:
    - Svet nie je čiernobiely, preto:
    - Všetko je sivé, preto:
    - Nikto nie je lepší než druhí, preto:
    - Môžem robiť čokoľvek sa mi zachce a vy ma nemôžete zastaviť, hahahaha.
- My však potrebujete vedieť zastaviť ľudí, aby nevraždili.
- Preto musí existovať nejaký spôsob, ako si byť absolútne istí, inak morálni relativisti vyhrajú.

Obrátená hlúposť nie je inteligencia. Nemôžete dôjsť k správnej odpovedi tým, že otočíte *každý jeden* riadok argumentu, ktorý viedol k nesprávnemu záveru – to dáva bláznovi príliš podrobnú kontrolu nad vami. Každý jeden riadok musí byť správny, aby matematický argument platil. A neplatí, že keď morálni relativisti povedia: „Svet nie je čiernobiely," musí to byť nepravda, rovnako ako neplatí, že ak si Stalin myslel, že 2 + 2 = 4, musí „2 + 2 = 4" byť nepravda. Chyba (a stačí, ak urobíte jednu) je v skoku z dvojfarebného pohľadu na jednofarebný, že všetky sivé farby majú rovnaký odtieň.

Bol by to príliš veľký ústupok (vlastne ústupok vo všetkom) súhlasiť s predpokladom, že musíte mať absolútne poznanie absolútne dobrých možností a absolútne zlých možností, aby ste mohli byť morálni. Môžete mať neisté poznanie relatívne lepších a relatívne horších možností, a aj tak si vyberať. V skutočnosti by to mala byť rutina, nie niečo, z čoho robíme drámu.

Myslím tým, áno, keby ste si museli vybrať medzi dvoma alternatívami A a B, a keby sa vám nejako podarilo získať určite správnu dobre kalibrovanú 100%-nú istotu, že A je absolútne a úplne žiadúce, a že B je suma všetkého zlého a odporného, bola by to *postačujúca* podmienka na výber A namiesto B. Ale nie je to *nevyhnutná* podmienka.

Aha, a: Logický klam: Odvolávanie sa na dôsledky názoru.

Pozrime sa, čo ešte potrebujú vedieť? No, že existuje celá racionalistická kultúra, ktorá hovorí, že pochybnosti, otázky a priznanie si chyby nie sú hrozné hanebné veci.

Existuje tu celá tá predstava, že môžeme získavať informácie tým, že sa *na veci pozeráme*, namiesto toho, že nás o nich presviedčajú. Keď sa pozriete na veci pozornejšie, niekedy zistíte, že sú iné ako ste si mysleli na prvý pohľad; ale to neznamená, že vám príroda klame, alebo že by ste sa mali prestať pozerať.

Ďalej existuje pojem kalibrovanej istoty – že „pravdepodobnosť" nie je to isté ako malý ukazovateľ vo vašej hlave, ktorý meria emocionálnu pripútanosť k myšlienke. Je to skôr miera toho, ako často v praxi, v skutočnom živote, ľudia v istom stave presvedčenia hovoria veci, ktoré sú naozaj pravda. Ak vezmete sto ľudí a požiadate ich, aby napísali sto výrokov, ktorými sú si „absolútne istí", koľko z nich bude správnych? Nie sto.

Dokonca tvrdenia, ohľadom ktorých sú ľudia naozaj fanatickí, majú *omnoho menšiu* pravdepodobnosť byť správnymi než tvrdenia typu „Slnko je väčšie ako Mesiac", ktoré vyzerajú príliš všedne na to, aby sa nad nimi niekto vzrušoval. Pre každé tvrdenie, ktorým si je niekto „absolútne istý" pravdepodobne nájdete niekoho, kto si je „absolútne istý" jeho opakom, pretože takéto fanatické vyznania presvedčenia nevznikajú v neprítomnosti opozície. Takže tento malý ukazovateľ v hlavách ľudí, ktorý meria ich emocionálnu pripútanosť k názoru, sa nedá ľahko preložiť na kalibrovanú istotu – dokonca sa ani nespráva monotónne.

A čo sa týka „absolútnej istoty" - nuž, ak poviete, že niečo má pravdepodobnosť 99,9999 %, znamená to, že si myslíte, že by ste mohli vysloviť *milión* rovnako silných nezávislých tvrdení, *jedno za druhým*, čo by vám zabralo asi celý rok, a mýlili by ste sa v priemere jedenkrát. To je dosť neuveriteľné. (Je úžasné uvedomiť si, že túto úroveň istoty dokážeme dosiahnuť výrokom: „Nevyhráš v lotérii".) Nehovorme už teda o pravdepodobnosti 1,0. Keď si uvedomíte, že na fungovanie v živote

nepotrebujete pravdepodobnosť 1,0, uvedomíte si, aké absolútne smiešne je myslieť si, že by ste ľudským mozgom niekedy mohli dosiahnuť 1,0. Pravdepodobnosť 1,0 nie je iba istota, je to *nekonečná istota*.

Vlastne sa mi zdá, že aby sme sa vyhli nepochopeniu verejnosti, možno by vedci mali hovoriť: „Nie sme si NEKONEČNE istí“ namiesto „Nie sme si istí“. Pretože to druhé v bežnej reči naznačuje, že máte konkrétne dôvody pochybovať.

*→

53. Ako ma presvedčiť, že 2 + 2 = 3

V kapitole „Čo je indícia?“ som napísal:

> To je dôvod, prečo racionalisti kladú taký silný dôraz na paradoxne vyzerajúce tvrdenie, že názor je skutočne *hodnotný* iba vtedy, keby bolo principiálne možné presvedčiť vás, aby ste si mysleli niečo iné. Keby vaša sietnica skončila v rovnakom stave bez ohľadu na to, aké svetlo na ňu dopadá, boli by ste slepí... Preto máme frázu „slepá viera“. Ak to, čomu veríte, nezávisí na tom, čo vidíte, potom vás oslepili prakticky rovnako, ako keby vám vypichli oči.

Cihan Baran odpovedal:

> Neviem si predstaviť situáciu, ktorá by urobila 2 + 2 = 4 nepravdivým. Asi z toho dôvodu je moje presvedčenie, že 2 + 2 = 4, nepodmienené.

Priznávam sa, že si neviem predstaviť „situáciu“, ktorá by urobila 2 + 2 = 4 nepravdivým. (Existujú redefinície, ale to nie sú „situácie“, a nehovorí sa v nich o 2, 4, =, ani +.) Ale to nerobí moje presvedčenie nepodmieneným. Viem si pomerne ľahko predstaviť situáciu, ktorá by ma *presvedčila*, že 2 + 2 = 3.

Predstavme si, že by som jedného rána vstal, vybral si dva štuple z uší a položil ich vedľa ďalších dvoch na mojom nočnom stolíku, a všimol by si, že teraz sú tam tri štuple, pričom žiaden štupeľ sa nezjavil ani nezmizol – v rozpore s mojou uloženou spomienkou, že 2 + 2 sa má rovnať 4. Navyše, keby som si predstavil celý ten proces vo svojej mysli, vyzeralo by to, že aby z XX a XX vzniklo XXXX, muselo by tam nejaké X pribudnúť odnikiaľ, a navyše by to nesedelo so zvyšnou matematikou, ktorú by som si predstavoval, lebo odobrať XX z XXX by dávalo XX, ale odobrať XX z XXXX by dávalo XXX. To by bolo v rozpore s mojou uloženou spomienkou, že 3 – 2 = 1, ale tá spomienka by bola absurdná zoči-voči fyzikálnemu a mentálnemu potvrdeniu, že XXX – XX = XX.

Ešte by som si to skontroloval na vreckovej kalkulačke, Googli, a povedzme mojej kópii knihy 1984, kde Winston píše, že „Sloboda znamená slobodu povedať, že dva plus dva sa rovná tri.“ Toto všetko by mi prirodzene ukázalo, že zvyšok sveta súhlasí s mojimi terajšími predstavami a nesúhlasí s mojou spomienkou, že 2 + 2 = 3.

Ako som mohol byť taký pomýlený, že som veril, že 2 + 2 = 4? Napadajú mi dve vysvetlenia: Po prvé, neurologická porucha (možno spôsobená kýchnutím) spôsobila, že sa všetky moje spomienky na sčítanie posunuli o 1 nahor. Po druhé, niekto sa so mnou zahráva, či už pomocou hypnózy alebo počítačovej simulácie. V tom druhom prípade by som si myslel, že je pravdepodobnejšie, že sa dotyčný pohral s mojimi matematickými *spomienkami*, než že by sa 2 + 2 naozaj *rovnalo* 4. Žiadne z týchto prijateľne znejúcich vysvetlení by mi nezabránilo všimnúť si, že som veľmi, veľmi, *veľmi* zmätený.

Čo by ma presvedčilo, že 2 + 2 = 3 je inými slovami celkom rovnaký druh indície ako ten, čo ma momentálne presviedča, že 2 + 2 = 4: Krížová paľba indícií z fyzického pozorovania, mentálnej vizualizácie a spoločenského súhlasu.

Boli časy, keď som nemal poňatia o tom, že 2 + 2 = 4. Nedošiel som k tomuto *novému* názoru náhodnými procesmi – potom by neexistoval konkrétny dôvod, prečo by si môj mozog uložil „2 + 2 = 4“ namiesto „2 + 2 = 7“. Fakt, že si môj mozog uložil odpoveď, ktorá sa prekvapivo podobá na to, čo sa stane, keď položím dva štuple do uší vedľa dvoch štupľov, si žiada vysvetlenie, aké previazanie vytvára toto zvláštne zrkadlenie mysle a skutočnosti.

→ http://lesswrong.com/lw/mn/absolute_authority/

Naozaj sú iba dve možnosti ohľadom názoru o fakte – buď sa tam ten názor dostal pomocou procesu, ktorý zväzuje myseľ so skutočnosťou, alebo nie. Ak nie, potom ten názor nemôže byť správny; nanajvýš ak zhodou okolností. Ak má ten názor trochu vnútornej zložitosti (vyžaduje si simulovanie počítačovým programom s dĺžkou viac než 10 bitov), priestor možností je dosť veľký na to, aby zhoda okolností zmizla.

Nepodmienené fakty nie sú to isté ako nepodmienené názory. Ak ma zviazaná indícia presvedčila, že fakt je nepodmienený, to neznamená, že som tomuto faktu vždy veril a nepotreboval som zviazanú indíciu.

Verím, že 2 + 2 = 4 a pripadá mi pomerne ľahké predstaviť si situáciu, ktorá by ma presvedčila, že 2 + 2 = 3. Konkrétne, ten istý druh situácie, aká má v súčasnosti presviedča, že 2 + 2 = 4. Nemusím sa teda báť, že som obeťou slepej viery.

Ak sú v obecenstve nejakí kresťania, *ktorí poznajú Bayesovu vetu* (žiadni numerofóbovia, prosím), môžem sa vás opýtať, aká situácia by vás presvedčila o pravdivosti islamu? Pravdepodobne by to bol ten istý druh situácie, ktorý je kauzálne zodpovedný za vašu terajšiu vieru v kresťanstvo: Keby vás vytiahli kričiacich z maternice moslimskej ženy, a keby vás vychovali rodičia, ktorí vám stále hovorili, že je dobré bezpodmienečne veriť v islam. Alebo je toho viac? Ak áno, aká situácia by vás presvedčila o islame, alebo prinajmenšom o nie-kresťanstve?

*→

54. Nekonečná istota

V kapitole Absolútna autorita som tvrdil, že *nepotrebujete* nekonečnú istotu.

> Keby ste si museli vybrať medzi dvoma alternatívami A a B, a keby sa vám nejako podarilo získať určite správnu dobre kalibrovanú 100%-nú istotu, že A je absolútne a úplne žiadúce, a že B je suma všetkého zlého a odporného, bola by to *postačujúca* podmienka na výber A namiesto B. Nie je to *nevyhnutná* podmienka... Môžete mať neisté poznanie relatívne lepších a relatívne horších možností, a aj tak si vyberať. V skutočnosti by to mala byť rutina.

V prípade výroku 2 + 2 = 4 musíme rozlišovať medzi mapou a územím. Vzhľadom na napohľad absolútnu stabilitu a všeobecnosť fyzikálnych zákonov je možné, že nikdy, v celej histórii vesmíru, žiadna častica neprekročila lokálnu hranicu rýchlosti svetla. To znamená, že hranica rýchlosti svetla môže byť pravdivá nielen 99 % času, alebo 99,9999 % času, alebo $1-1/10^{10^{100}}$ času, ale jednoducho *vždy a absolútne pravdivá*.

Ale či my môžeme niekedy mať *absolútnu istotu* v hranicu rýchlosti svetla, to je úplne iná otázka. Mapa nie je územie.

Môže byť úplná pravda, že nejaký študent opísal svoju úlohu, ale či máte o tomto fakte nejaké vedomosti – tobôž *absolútnu* istotu v tento názor – je samostatná téma. Ak si hodíte mincou a nepozriete na ňu, môže byť úplná pravda, že na minci padla hlava, a vy si môžete byť úplne neistí, či na minci padla hlava alebo znak. Stupeň neistoty nie je to isté ako stupeň pravdy alebo frekvencia výskytu.

To isté platí pre matematické pravdy. Je otázka, či možno výrok „2 + 2 = 4“ alebo „V Peanovej aritmetike SS0 + SS0 = SSSS0“ označiť za *pravdivý* v čisto abstraktnom zmysle, bez ohľadu na fyzické systémy, ktoré vyzerajú, že sa správajú v súlade s Peanovými axiómami. Ale keď už som toto povedal, pôjdem na vec a tipnem si, že v ľubovoľnom zmysle, v ktorom je „2 + 2 = 4“ pravda, je to vždy a presná pravda, nie iba približne pravda („2 + 2 sa naozaj rovná 4,000 000 4“) ani pravda v 999 999 999 999 prípadoch z 1 000 000 000 000.

Nie som si celom istým, čo v tomto prípade „pravda“ znamená, ale stojím za svojím odhadom. Dôveryhodnosť výroku „2 + 2 = 4 je vždy pravda“ je omnoho väčšia než dôveryhodnosť ľubovoľného filozofického názoru na to, čo presne znamená „pravda“, „vždy“ a „je“ v uvedenom výroku.

→ http://lesswrong.com/lw/jr/how_to_convince_me_that_2_2_3/

To však neznamená, že mám *absolútnu istotu*, že 2 + 2 = 4. Pozrite predchádzajúcu diskusiu o tom, ako ma presvedčiť, že 2 + 2 = 3, čo možno dosiahnuť použitím rovnakých indícií ako tie, čo ma pôvodne presvedčili, že 2 + 2 = 4. Možno všetky moje doterajšie skúsenosti boli halucinácie alebo si ich nesprávne pamätám. Neurológovia poznajú aj čudnejšie poruchy mozgu než toto.

Ak teda pripíšeme nejakú pravdepodobnosť výroku „2 + 2 = 4“, aká by to mala byť pravdepodobnosť? V prípadoch ako je tento hľadáme dobrú kalibráciu – aby sa vety, ktorým pripíšete „pravdepodobnosť 99 %“ ukázali ako pravdivé v 99 prípadoch zo 100. Toto je v skutočnosti omnoho zložitejšie než si myslíte. Vezmite si sto ľudí a požiadajte každého z nich, aby povedal desať výrokov, ktorými si je „istý na 99 %“. Myslíte si, že z týchto 1000 výrokov bude nesprávnych približne 10?

Nebudem teraz diskutovať skutočné pokusy, ktoré sa robili s kalibráciou – nájdete ich v kapitole inej mojej knihy, „Kognitívne skreslenia potenciálne ovplyvňujúce hodnotenie globálnych rizík“ – pretože som videl, že keď to len tak vyhŕknem na nepripravených ľudí, začnú to používať ako Úplne Všeobecný Protiargument, ktorý im akosi naskočí do hlavy vždy, keď potrebujú znevážiť istotu niekoho, koho názor sa im nepáči, ale zabudnú si na to spomenúť, keď uvažujú o svojich vlastných názoroch. Pokúšam sa teda nehovoriť o pokusoch s kalibráciou, jedine ako súčasť štrukturovanej prednášky o rozumnosti, ktorá zahŕňa varovanie pred motivovaným skepticizmom.

Lenže pozorovaná kalibrácia ľudí, ktorí hovoria, že sú si „na 99 % istí“ nie je presnosť 99 %.

Predstavte si, že poviete, že ste si na 99,99 % istí, že 2 + 2 = 4. Tým tvrdíte, že by ste mohli urobiť 10 000 *nezávislých* výrokov, v ktoré máte rovnakú dôveru, a mýlili by ste sa v priemere jedenkrát. Možno pre 2 + 2 = 4 je tento mimoriadny stupeň istoty možný: „2 + 2 = 4“ je extrémne jednoduché, matematicky aj empiricky, a všeobecne spoločensky prijímané (nie s vášnivým potvrdením, ale ticho považované za samozrejmosť). Možno tu naozaj môžete mať istotu 99,99 %.

Nemyslím si, že môžete mať istotu 99,99 % pri výrokoch ako je „53 je prvočíslo“. Áno, vyzerá to pravdepodobne, ale dokiaľ by ste si vytvorili protokoly, ktoré vám umožnia tvrdiť 10 000 *nezávislých* výrokov tohto druhu – čiže nie iba sadu výrokov o prvočíslach, ale zakaždým nový protokol – pomýlili by ste sa viac než iba raz. Peter de Blanc má na túto tému zábavnú anekdotu. (Povedal som mu, aby to viac nerobil.)

Lenže mapa nie je územie: ak poviem, že som si na 99 % istý, že 2 + 2 = 4, neznamená to, že si myslím, že „2 + 2 = 4“ je pravda s presnosťou 99 %, alebo že „2 + 2 = 4“ je pravda v 99 prípadoch zo 100. Výrok, ktorému pripisujem takúto istotu je výrok, že „2 + 2 = 4 vždy a presne“, nie výrok „2 + 2 = 4 približne a väčšinou“.

A čo sa týka predstavy, že by ste mohli mať istotu až 100 % o matematickom výroku – no, teraz vážne! Keby ste povedali istotu 99,999 9 %, tvrdili by ste, že by ste mohli urobiť *milión* rovnako dôveryhodných výrokov, jeden za druhým, a mýlili by ste sa v priemere jedenkrát. To by zabralo asi jeden celý rok rozprávania, keby ste povedali jeden výrok každých 20 sekúnd a rozprávali by ste 16 hodín denne.

Ak tvrdíte istotu 99,999 999 999 9 %, dvíhate to na bilión. Bude teraz rozprávať sto ľudských životov a nepomýlite sa pritom ani raz?

Ak tvrdíte istotu $1-1/10^{10^{100}}$, vaše ego výrazne prevyšuje ego duševne chorých pacientov, ktorí sa považujú za bohov.

Pritom $1/10^{10^{100}}$ je omnoho menšie než aj tie pomerne malé nepredstaviteľne veľké čísla ako je $3\uparrow\uparrow\uparrow 3$. Ale ani istota $1-1/3\uparrow\uparrow\uparrow 3$ nie je až o toľko bližšia k **PRAVDEPODOBNOSTI 1** ako istota 90 %.

Keby všetko ostatné zlyhalo, hypotetickí Temní páni Matrixu, ktorí *práve teraz* zasahujú do toho, ako váš mozog pripisuje dôveryhodnosť *samotnej tejto vete*, nám zahatajú cestu a ušetria nás pliagy nekonečnej istoty.

Som si tým absolútne istý?

Samozrejme, že nie.

Ako raz povedal Rafal Smigrodski:

Povedal by som, že by sme mali dokázať priradiť matematickým pojmom nevyhnutným na odvodenie Bayesovho pravidla istotu menšiu než 1, a aj tak ho prakticky používať. Nie som si celkom istý, že si vždy musím byť neistý. Možno si môžem byť o niektorých veciach legitímne istý. Ale keď raz nejakému výroku priradím pravdepodobnosť 1, už to neviem nikdy zrušiť. Bez ohľadu na to, čo uvidím alebo čo sa naučím, budem musieť odmietnuť všetko, čo odporuje tejto axióme. Nepáči sa mi predstava, že by som už nikdy nedokázal zmeniť názor.

*→

55. 0 a 1 nie sú pravdepodobnosti

1, 2 a 3 sú celé čísla, takisto aj -4. Ak budete ďalej počítať nahor alebo počítať nadol, isto stretnete mnoho ďalších celých čísel. Nestretnete však nič, čo by sa volalo „kladné nekonečno“ alebo „záporné nekonečno“, takže tie nie sú celé čísla.

Kladné a záporné nekonečno nie sú celé čísla, ale skôr špeciálne symboly na rozprávanie o správaní celých čísel. Ľudia niekedy hovoria veci ako „5 + nekonečno = nekonečno“, pretože ak začnete pri 5 a budete počítať smerom nahor bez zastavenia, budete dostávať väčšie a väčšie čísla bez ohraničenia. Ale z toho nevyplýva, že „nekonečno – nekonečno = 5“. Nemôžete začať počítať od 0 nahor bez zastavenia, a potom počítať nadol bez zastavenia, a nakoniec zistiť, že ste na čísle 5.

Z tohto vidíme, že nekonečno nielenže nie je celé číslo, ale ani sa *nespráva* ako celé číslo. Ak sa nemúdro pokúsite zmiešať nekonečná s celými číslami, budete potrebovať všakovaké špeciálne nové napohľad nekonzistentné správania, ktoré ste nepotrebovali pre 1, 2, 3 a ďalšie *skutočné* celé čísla.

Aj keď nekonečno nie je celé číslo, nemusíte sa báť toho, že budete mať nedostatok čísel. Ľudia už videli päť oviec, milión zrniek piesku, a septilióny atómov, nikto však nikdy nespočítal nekonečne veľa niečoho. To isté platí pre spojité množstvá – ľudia odmerali milimetre veľké zrnká prachu, metre veľké zvieratá, kilometre veľké mestá, tisíce svetelných rokov veľké galaxie, ale nikto ešte nenameral niečo nekonečne veľké. V skutočnom svete nekonečná veľmi *nepotrebujete*.

(Poznámka pre sofistikovanejších čitateľov: Nemusíte mi písať zložité vysvetlenia rozdielov medzi povedzme ordinálnymi a kardinálnymi číslami. Áno, poznám rôzne pokročilé definície nekonečna z teórie množín, ale nevidím pre ne dobré použitie v teórii pravdepodobnosti. Viď nižšie.)

Pri zvyčajnom spôsobe zápisu pravdepodobností sú pravdepodobnosti medzi 0 a 1. Minca môže mať pravdepodobnosť 0,5, že na nej padne znak, alebo hlásateľ priradí pravdepodobnosť 0,9, že bude zajtra pršať.

To však nie je jediný možný zápis pravdepodobností. Môžete napríklad transformovať pravdepodobnosti na šance pomocou transformácie Š = (P / (1 – P)). Pravdepodobnosť 50 % by teda bola šancou 0,5 / 0,5 alebo 1, čo zvyčajne zapisujeme 1 : 1, zatiaľ čo pravdepodobnosť 0,9 by bola šancou 0,9 / 0,1, čo zvyčajne zapisujeme 9 : 1. Aby ste zo šancí urobili naspäť pravdepodobnosti, použijete P = (Š / (1 + Š)) a je to dokonale reverzibilné, takže táto transformácia je izomorfizmus – obojstranné reverzibilné mapovanie. Pravdepodobnosti a šance sú teda izomorfné, a môžete používať jedno alebo druhé podľa toho, čo je pohodlnejšie.

Je napríklad pohodlnejšie používať šance, keď robíte bayesovské aktualizácie. Povedzme, že hodím šesťstennou kockou: Ak padne iné číslo ako 1, je šanca 10 %, že začujem zvonenie, ale ak padne 1, je šanca 20 %, že začujem zvonenie. Hodil som kockou a počul som zvonenie. Aká je *šanca*, že padlo číslo 1? Nuž, počiatočná šanca je 1 : 5 (čo zodpovedá reálnemu číslu 1 / 5 = 0,20) a pomer podmienených pravdepodobností je 0,2 : 0,1 (čo zodpovedá reálnemu číslu 2) a tieto môžem jednoducho navzájom vynásobiť a dostanem výslednú šancu 2 : 5 (čo zodpovedá reálnemu číslu 2 / 5 čiže 0,40). Potom to premením späť na pravdepodobnosti, ak chcem, a dostanem (0,4 / 1,4) = 2 / 7 = ~29%.

→ http://lesswrong.com/lw/mo/infinite_certainty/

So šancami sa teda pri bayesovských aktualizáciách lepšie pracuje – ak používate pravdepodobnosti, musíte používať Bayesovu vetu v jej zložitejšom tvare. Ale pravdepodobnosti sa lepšie hodia na odpovedanie na otázky ako: „Ak hodím šesťstennou kockou, aká je šanca, že uvidím číslo od 1 do 4?“ Môžete sčítať pravdepodobnosti 1 / 6 pre každú stranu a dostať 4 / 6, nemôžete však sčítať šance 0,2 pre každú stranu a dostať šancu 0,8.

Prečo toto všetko hovorím? Aby som ukázal, že „pomery šancí“ sú rovnako legitímny spôsob mapovania neistoty na reálne čísla ako „pravdepodobnosti“. Pri niektorých operáciách sú pohodlnejšie pomery šancí, pri iných pravdepodobnosti. Slávny dôkaz nazývaný Coxova veta (plus jej rôzne rozšírenia a upresnenia) ukazuje, že všetky spôsoby reprezentovanie neistoty, ktoré dodržiavajú určité rozumne znejúce obmedzenia, sú v konečnom dôsledku izomorfné.

Prečo záleží na tom, či sú pomery šancí rovnako legitímne ako pravdepodobnosti? Pravdepodobnosti sa štandardne zapisujú medzi 0 a 1, pričom aj 0 aj 1 vyzerajú, že by to mali byť pomerne dosiahnuteľné hodnoty – je ľahké vidieť 1 zebru alebo 0 jednorožcov. Keď si však transformujete pravdepodobnosti na pomery šancí, 0 zostane 0, ale 1 ide do kladného nekonečna. Teraz absolútna pravda nevyzerá až tak ľahko dosiahnuteľná.

Reprezentácia, v ktorej sa bayesovské aktualizácie robia ešte ľahšie, je logaritmus šance – takto E. T. Jaynes odporúčal myslieť na pravdepodobnosti. Povedzme napríklad, že pôvodná pravdepodobnosť nejakého tvrdenia je 0,0001 – to zodpovedá logaritmu šance okolo -40 decibelov. Potom vidíte indíciu, ktoré je 100-krát pravdepodobnejšia, ak je tvrdenie pravdivé, než ak je nepravdivé. To je 20 decibelov indície. Výsledný pravdepodobnosť je teda zhruba -40 db + 20 db = -20 db, čo znamená, že výsledná pravdepodobnosť je ~0,01.

Keď transformujete pravdepodobnosti na logaritmy šancí, z 0 sa stane záporné nekonečno a z 1 sa stane kladné nekonečno. Teraz aj nekonečná istota aj nekonečná nepravdepodobnosť vyzerajú dosť nedosiahnuteľne.

Pri pravdepodobnostiach sa zdá, že medzi 0,999 9 a 0,999 99 je rozdiel iba 0,000 09, takže 0,502 je omnoho ďalej od 0,503 než je 0,999 9 od 0,999 99. Aby sme z pravdepodobnosti 0,999 99 dostali pravdepodobnosť 1, vyzerá to, že treba prekonať iba vzdialenosť 0,000 01.

Ale keď to transformujeme na pomery šancí, z 0,502 a 0,503 sa stane 1,008 a 1,012, a z 0,999 9 a 0,999 99 sa stane 9,999 a 99,999. A keď to transformujeme na logaritmy šancí, z 0,502 a 0,503 sa stane 0,03 decibelov a 0,05 decibelov, ale z 0,999 9 a 0,999 99 sa stane 40 decibelov a 50 decibelov.

Keď pracujete s logaritmami šancí, **vzdialenosť medzi ľubovoľnými dvoma stupňami neistoty sa rovná množstvu indície, ktoré by ste potrebovali, aby ste sa dostali z jedného na druhý**. To znamená, že logaritmy šancí nám dávajú prirodzenú mieru vzdialenosti medzi stupňami istoty.

Použitie logaritmov šancí odhaľuje fakt, že dosiahnutie nekonečnej istoty si vyžaduje nekonečne silnú indíciu, rovnako ako si nekonečná absurdita vyžaduje nekonečne silnú protiindíciu.

Ďalej, všetky druhy štandardných viet o pravdepodobnosti majú špeciálne prípady, ak sa do nich pokúsite vložiť 1-y alebo 0-y – ako sa napríklad stane, ak skúšate robiť bayesovskú aktualizáciu po pozorovaní, ktorému ste priradili pravdepodobnosť 0.

Navrhujem teda, že by dávalo zmysel povedať, že 1 a 0 nie sú pravdepodobnosti; rovnako ako záporné a kladné nekonečno, ktoré nespĺňajú axiómy poľa, nie sú medzi reálnymi číslami.

Hlavný dôvod, prečo by toto mohlo znepokojiť teoretikov pravdepodobnosti je, že by sme museli znovu odvodiť vety, ktoré sme predtým získali z predpokladu, že môžeme zjednotenie pravdepodobností vybaviť tým, že sčítame všetky kúsky a povieme, že ich súčet je 1.

Lenže v skutočnom svete, keď hodíte kocku, nemáte doslova nekonečnú istotu, že padne nejaké číslo od 1 do 6. Kocka môže zastať na hrane; môže na ňu padnúť meteor; alebo zasiahnu Temní páni Matrixu a napíšu „37“ na jednu stranu.

Keby ste si vyrobili čarovný symbol, ktorý by znamenal „všetky pravdepodobnosti, ktoré som nezvážil“, mohli by ste sčítať všetky udalosti vrátane tohto čarovného symbolu a získali by ste čarovný symbol „T“, ktorý by znamenal nekonečnú istotu.

Ale ja by som sa radšej opýtal, či existuje nejaký spôsob, ako odvodiť vety bez použitia čarovných symbolov so špeciálnym správaním. To by bolo elegantnejšie. Rovnako ako existujú matematici, ktorí odmietajú uveriť v dvojitú negáciu alebo nekonečné množiny, ja by som rád bol teoretikom pravdepodobnosti, ktorý neverí v nekonečnú istotu.

*→

56. Záleží mi na vašej rozumnosti

Niektoré reakcie na článok Lotérie: Plytvanie nádejou ma karhali za opovážlivosť kritizovať rozhodnutia druhých; ak sa niekto iný rozhodne kúpiť si žreb lotérie, kto som ja, aby som nesúhlasil? Toto je špeciálny prípad všeobecnejšej otázky: Čo ma je do toho, ak sa niekto iný rozhodne veriť tomu, čo je príjemné, namiesto toho, čo je pravdivé? Nemôže si každý vybrať sám pre seba, či mu záleží na pravde?

Dá sa ľahko odvrknúť: „A čo je *teba* do toho, či *mne* záleží na tom, či niekomu *inému* záleží na pravde?“ Je trochu nekonzistentné, ak vaša funkcia úžitku obsahuje negatívne znamienko pre to, keď funkcia úžitku niekoho iného obsahuje znamienko pre funkciu úžitku niekoho iného. Ale to je iba odvrknutie, nie odpoveď.

Tu je teda moja odpoveď: Verím tomu, že je správne, aby som ako človek mal záujem o budúcnosť a o to, čím sa ľudská civilizácia v budúcnosti stane. Tento záujem sa týka aj ľudského hľadania pravdy, ktoré postupom generácií silnelo (lebo nie vždy existovala veda). Chcem, aby toto hľadanie ešte zosilnelo v *tejto* generácii. To je moje želanie ohľadom budúcnosti. Lebo my všetci sme hráčmi na obrovskej hracej ploche, bez ohľadu na to, či túto zodpovednosť prijímame alebo nie.

Preto *mi* záleží na *vašej* rozumnosti.

Je toto nebezpečná myšlienka? Áno, a nielen trošku príjemne „nebezpečná“. Ľudia už boli upálení zaživa preto, lebo sa nejaký kňaz rozhodol, že nerozmýšľali tak, ako by mali. Rozhodnutie upaľovať ľudí zaživa pretože „nerozmýšľajú správne“ - to je odporný druh uvažovania, však? Nechcete, aby ľudia uvažovali takto, načo, je to *hnusné*. Ľudia, ktorí takto rozmýšľajú, nuž, budeme s nimi musieť niečo urobiť...

Súhlasím! Tu je môj návrh: Nech sa argumentuje proti zlým myšlienkam, ale *nehádžme* ich nositeľov do ohňa.

Záver, ktorému sa túžime vyhnúť, znie: „Myslím si, že Susie povedala niečo zlé, *preto* treba Susie upáliť.“ Niektorí sa tomuto záveru snažia vyhnúť tak, že označia za neprípustné myslieť si, že Susie povedala niečo zlé. Nikto by nikoho nemal súdiť, nikdy; každý, kto súdi, sa dopúšťa strašného hriechu a mal by zaň byť verejne pranierovaný.

Čo sa mňa týka, nesúhlasím s *druhou* časťou tvrdenia. Môj záver je: „Myslím si, že Susie povedala niečo zlé, *preto* budem argumentovať proti tomu, čo povedala, ale nebudem ju hádzať do ohňa, ani sa jej pokúšať brániť v rozprávaní násilím či reguláciou...“

Všetci sme hráčmi na obrovskej hracej ploche, a časťou môjho záujmu o budúcnosť je urobiť túto hru férovou. Nesamozrejmá myšlienka, na ktorej je postavená veda, znie, že nezhody ohľadom faktov by sa mali riešiť pomocou experimentov a matematiky, nie násilia a ediktov. Tento neuveriteľný prístup sa dá rozšíriť aj mimo vedu, do férového zápasu o celú budúcnosť. Mali by ste mať povinnosť vyhrať tým, že presvedčíte ľudí, a nemali by ste mať dovolené upaľovať ich. Toto je jeden z princípov rozumnosti, ktorým som prisahal vernosť.

Ľudia, ktorí obhajujú relativizmus alebo sebectvo mi nepripadajú byť naozaj relativistická alebo sebecký. Keby boli naozaj relativistickí, neodsudzovali by nič. Keby boli naozaj sebeckí, radšej by šli zarábať peniaze než vášnivo argumentovať s druhými. Skôr si len vybrali hrať za stranu relativizmu, ktorej cieľom na tejto obrovskej hracej ploche je zabrániť hráčom – *všetkým* – robiť určitý druh úsudku.

→ http://lesswrong.com/lw/mp/0_and_1_are_not_probabilities/

Alebo si vybrali stranu sebectva, ktorej cieľom je urobiť *všetkých* hráčov sebcami. A potom hrajú túto hru, v súlade alebo nesúlade so svojou múdrosťou.

Ak existujú aj nejakí skutoční relativisti alebo sebci, nepočujeme ich – zostávajú tichí, nehrajú.

Nemôžem si pomôcť, záleží mi na tom, ako rozmýšľate, pretože – nemôžem si pomôcť, ale takto vidím vesmír – vždy keď sa nejaký človek odvráti od pravdy, rozvíjajúci sa príbeh ľudstva sa stane o čosi temnejším. V mnohých prípadoch je to len maličká temnota. (Nie *vždy* to skončí ubližovaním.) Klamať sám sebe, v súkromí svojich myšlienok, nezatieňuje históriu ľudstva natoľko ako klamanie verejnosti alebo upaľovanie ľudí. Napriek tomu, čosi vo mne smúti aj nad tým. A pokiaľ sa vás *nepokúšam* hodiť do ohňa – iba argumentujem proti vašim myšlienkam – verím, že je to správne pre mňa, ako človeka, ktorému záleží na jeho blížnych. A toto je aj pozícia, ktorú bránim pre budúcnosť.

*→

→ http://lesswrong.com/lw/hn/your_rationality_is_my_business/

F: Politika a rozumnosť

57. Politika zabíja myslenie

Ľuďom sa v hlavách dejú smiešne veci, keď hovoria o politike. Evolučné dôvody sú také zrejme, že sa ich oplatí spomenúť: V pravekom prostredí bola politika otázkou života a smrti. A sexu a bohatstva a spojencov a povesti... Keď sa dnes dostanete do hádky, či by „sme“ mali zvýšiť minimálnu mzdu, vykonávate adaptáciu na praveké prostredie, kde byť na nesprávnej strane hádky mohlo znamenať smrť. Byť na *správnej* strane hádky mohlo *vám* umožniť zabiť svojho nenávideného soka!

Ak chcete niečo povedať o vede alebo rozumnosti, odporúčam nevyberať si príklady zo *súčasnej* politiky, ak sa tomu dá vyhnúť. Ak vaša pointa neodmysliteľne súvisí s politikou, potom hovorte o Ľudovítovi XVI. počas francúzskej revolúcie. Politika je dôležitá oblasť, v ktorej by sme jednotlivo mali používať svoju rozumnosť – ale je to otrasná oblasť na *učenie sa* rozumnosti alebo na diskutovanie o rozumnosti, okrem prípadu, že všetci diskutujúci už rozumní sú.

Politika je pokračovanie vojny inými prostriedkami. Argumenty sú vojaci. Akonáhle viete, na ktorej strane ste, musíte podporovať všetky argumenty tejto strany a napádať všetky argumenty, ktoré vyzerajú v prospech nepriateľskej strany; inak je to, akoby ste vlastných vojakov bodali do chrbta a poskytovali pomoc a podporu nepriateľom. Ľudia, ktorí by s chladnou hlavou spravodlivo zvážili zo všetkých strán nejakú tému vo svojom pracovnom živote ako vedci, sa dokážu náhle premeniť na zombie skandujúce heslá, ak k danej téme existuje Modrý alebo Zelený postoj.

V oblasti umelej inteligencie, špeciálne v oblasti nemonotónneho uvažovania, existuje štandardný problém: „Všetci kvakeri sú pacifisti. Žiaden republikán nie je pacifista. Nixon je kvaker a republikán. Je Nixon pacifista?“

Aký zmysel to prosím vás malo, vybrať takýto príklad? Vyburcovať politické emócie čitateľov a odpútať ich pozornosť od hlavnej otázky? Dosiahnuť, aby sa republikáni cítili nevítaní na kurzoch umelej inteligencie a odradiť ich od výberu takéhoto smeru? (Nie, skôr než sa na to niekto opýta, nie som republikán. Ani demokrat.)

Prečo by si niekto vybral takýto *rozptyľujúci* príklad na ilustráciu nemonotónneho uvažovania? Autor pravdepodobne nedokázal odolať pokušeniu dobre si štuchnúť do tých nenávidených Zelených. Je to taký *dobrý* pocit uštedriť poriadny úder, viete, to je ako keby ste skúšali odolať čokoládovému koláču.

Podobne ako s čokoládovým koláčom, nie všetko, čo príjemne chutí, je aj dobré.

Nehovorím, že by sme mali byť apolitickí, dokonca ani že by sme mali prijať ideál neutrálneho uhla pohľadu z Wikipédie. Pokúsme sa však odolávať pokušeniu týchto dobrých štuchancov, ak sa tomu môžeme vyhnúť. Ak vaša téma legitímne súvisí s pokusmi zakázať evolúciu zo školských osnov, potom nech sa páči, hovorte o tom – ale neobviňujte z toho explicitne celú Republikánsku stranu; niektorí z vašich čitateľov môžu byť republikáni, no môžu mať pocit, že problémom je pár fanatikov, nie celá strana. Podobne ako pri neutrálnom pohľade Wikipédie, nie je podstatné, či (si myslíte, že) na vine naozaj *je* Republikánska strana. Pre duchovný rast komunity je skrátka lepšie diskutovať o téme bez privolávania politických farieb.

⁂→

→ http://www.lesswrong.com/lw/gw/politics_is_the_mindkiller/

58. Debaty o pravidlách by nemali vyzerať jednostranne

Robin Hanson nedávno navrhol obchody, v ktorých by sa mohli predávať zakázané výrobky. Existuje viacero výborných argumentov v prospech takéhoto zákona – jednotlivec má neodňateľné právo na slobodu, byrokrati majú kariérnu motiváciu zakazovať *všetko*, zákonodarcovia sú rovnako omylní ako občania. Napriek tomu (som napísal, že) do takéhoto obchodu príde aj *nejaká* chudobná, úprimná, neveľmi vzdelaná matka piatich detí a kúpi si „Nápoj z kyseliny sírovej Dr. Šarlatána“ proti artritíde a zomrie, zanechajúc siroty, ktoré budú plakať vo verejnoprávnej televízii.

Iba som konštatoval jednoduchý fakt. Prečo si niektorí ľudia mysleli, že to bol argument v prospech regulácie?

Pri jednoduchých faktických otázkach (napríklad či život na Zemi vznikol prirodzeným výberom) môžeme legitímne očakávať, že dokazovanie bude jednostrannou bitkou; samotné fakty sú buď tak alebo onak a takzvaná „vyvážená argumentácia“ by toto mala odrážať. Podľa bayesovskej definície indície je „silná indícia“ presne ten typ indície, o ktorom očakávame, že sa bude nachádzať iba na jednej strane dôkazu.

Nie je však dôvod, aby takúto jednostrannosť vykazovali aj zložité aktivity s mnohými dôsledkami. Prečo sa teda zdá, že ľudia chcú, aby ich debaty o *pravidlách* boli jednostranné?

Politika zabíja myslenie. Argumenty sú vojaci. Akonáhle viete, na ktorej strane ste, musíte podporovať všetky argumenty tejto strany a napádať všetky argumenty, ktoré vyzerajú v prospech nepriateľskej strany; inak je to, ako keby ste vlastných vojakov bodali do chrbta. Ak sa držíte tohto vzorca, budú aj vám debaty o pravidlách pripadať jednostranné – náklady a nevýhody vášho obľúbeného riešenia sú nepriateľskí vojaci, na ktorých treba útočiť všetkými možnými spôsobmi.

Človek by sa mal vyvarovať aj ďalšieho chybného vzorca, predstavy, že Hlboká Múdrosť si žiada robiť dokonale vyrovnané kompromisy medzi tými dvoma názormi na pravidlá, ktoré dostávajú najviac priestoru v médiách. Zákon môže legitímne mať *nevyvážené* náklady a prínosy. Keby politické otázky neboli naklonené jedným alebo druhým smerom, nedokázali by sme sa o nich rozhodovať. Je tu však aj ľudský sklon popierať všetky náklady obľúbeného riešenia alebo popierať všetky prínosy neobľúbeného riešenia; a ľudia si teda budú myslieť, že táto nerovnováha je naklonená omnoho viac, než v skutočnosti je.

Ak povolíte obchody, v ktorých sa predávajú inak zakázané produkty, nejaká chudobná, úprimná, neveľmi vzdelaná matka piatich detí si tam kúpi niečo, čo ju zabije. Toto je predpoveď faktického dôsledku a ako faktická otázka sa zdá byť celkom priamočiara – príčetný človek by mal byť pripravený uznať, že je to pravda, bez ohľadu na to, aký postoj má k danej politickej téme. Môžete si *tiež* myslieť, že postaviť veci mimo zákon akurát spôsobí, že budú drahšie, že regulátori zneužívajú svoje právomoci, alebo že jej sloboda ako jednotlivca prebíja vašu túžbu zasahovať do jej života. Avšak, vecou jednoduchého faktu je, že ona napriek tomu zomrie.

Žijeme v nespravodlivom vesmíre. Ako všetci primáti, aj ľudia majú silnú negatívnu reakciu na vnímanú nespravodlivosť; preto nás tento fakt stresuje. Existujú dve obľúbené metódy, ako si s výslednou kognitívnou disonanciou poradiť. Po prvé, človek môže zmeniť svoj pohľad na fakty – poprie, že sa daná nespravodlivá udalosť stala, alebo zmení históriu, aby vyzerala spravodlivejšie. (Sprostredkuje to afektívna heuristika a klam spravodlivého sveta.) Po druhé, človek môže zmeniť svoju morálku – poprieť, že daná udalosť bola nespravodlivá.

Niektorí libertariáni by mohli povedať, že keď pôjdete do „obchodu so zakázanými výrobkami“, prejdete okolo jasných varovných nápisov hovoriacich: „VECI V TOMTO OBCHODE VÁS MÔŽU ZABIŤ“ a kúpite si niečo, čo vás zabije, potom je to vaša vlastná chyba a zaslúžite si to. Ak by toto bola morálna pravda, potom by existencia obchodov predávajúcich zakázané výrobky nemala *žiadnu nevýhodu*. Nielenže by to bol *v konečnom dôsledku zisk*, ale bola by to *jednostranná* zmena bez akýchkoľvek nevýhod.

Iní argumentujú, že regulátorov možno naučiť, aby sa rozhodovali rozumne a v harmónii so záujmami spotrebiteľa; keby toto bola faktická pravda, potom (z ich morálneho pohľadu) by regulácia nemala *žiadnu nevýhodu*.

Či sa vám páči alebo nie, inteligenciu dostane človek náhodne pridelenú podľa narodenia – aj keď toto je jeden z prípadov, kde je nespravodlivosť vesmíru taká extrémna, že mnohí ľudia sa rozhodnú popierať fakty. Experimentálne indície pre čisto genetickú zložku 0,6 – 0,8 majú prevahu, ale aj keby ste toto popreli, rodičovskú výchovu alebo základnú školu si človek tiež nevyberá.

Ja som bol vychovaný, aby som veril, že popierať skutočnosť je *morálne nesprávne*. Keby som sa mal zapojiť do túžobného optimizmu, ako mi Nápoj z kyseliny sírovej pravdepodobne urobí lepšie, robil by som niečo, pred čím ma *varovali*, čo ma vychovali vnímať ako neprijateľné. Niektorí ľudia sa narodia do prostredia – neriešme teraz ich gény, pretože tá časť je príliš nespravodlivá – kde im miestny šaman povie, že je *správne* mať vieru a *nesprávne* byť skeptický. S najlepšou vôľou sa budú riadiť touto radou a zomrú. Na rozdiel od vás ich nevychovali, aby verili, že ľudia sú zodpovední za svoje individuálne rozhodnutie nasledovať smerovanie spoločnosti. Naozaj si myslíte, že vy sami ste takí chytrí, že by ste boli správne vedecky skeptickí, aj keby ste sa narodili v roku 500 nášho letopočtu? Áno, narodiť sa je lotéria, bez ohľadu na to, čo si myslíte o génoch.

Povedať: „Ľudia, ktorí si kupujú nebezpečné výrobky, si zaslúžia trpieť!“ nie je zásadovosť. Je to spôsob, ako popierať, že žijeme v nespravodlivom vesmíre. Skutočná zásadovosť je povedať: „Áno, kyselina sírová spôsobuje strašnú bolestivú smrť, a nie, tá matka piatich detí si to nezaslúžila, ale napriek tomu necháme tieto obchody otvorené, pretože sme urobili takýto výpočet nákladov a prínosov.“ Viete si predstaviť, že nejaký politik povie toto? Ja tiež nie. Ale keď majú ekonómovia možnosť ovplyvňovať politické rozhodnutia, pomohlo by, keby si to vedeli aspoň myslieť v súkromí – možno dokonca aj povedať v odborných článkoch, príslušne zamaskované mnohoslabičnou terminológiou, aby ich médiá nemohli citovať.

Nemyslím si, že ak niekto urobí hlúpu chybu a zomrie, že je to dôvod oslavovať. Ja to počítam za tragédiu. Nie vždy je dobré pomáhať ľuďom, zachraňovať ich pred dôsledkami vlastných činov; pri treste smrti však robím morálnu hranicu. Ak ste mŕtvi, nemôžete sa použiť z vlastných chýb.

Nanešťastie, tento vesmír so mnou nesúhlasí. Uvidíme, kto z nás zostane stáť ako víťaz, keď toto celé skončí.

*→

59. Váhy spravodlivosti, zápisník rozumnosti

Pani Spravodlivosť sa zvyčajne zobrazuje držiac váhy. Váhy majú tú vlastnosť, že keď hocičo potiahne jednu stranu nadol, posunie tým druhú stranu nahor. Vďaka tomu sa veci veľmi pohodlne a ľahko sledujú. Je to zvyčajne aj hrubé skreslenie.

Ako ľudia máme prirodzený sklon považovať diskusiu za istú formu súboja, pokračovanie vojny, šport; a v športe stačí sledovať, koľko bodov získalo ktoré družstvo. Sú iba dve družstvá a každý bod v neprospech jednej strany je bodom v prospech tej druhej. Každý v hľadisku si v duchu počíta, koľko bodov ktorý rečník získal voči druhému. Na konci diskusie je rečník s väčším bodovým ziskom, prirodzene, víťazom; všetko, čo povedal, musí byť teda pravda a všetko, čo povedal porazený, musí byť zle.

„Afektívna heuristika v posudzovaní rizika a úžitku“ skúmala, či pokusné osoby zmiešavajú svoje hodnotenia možných úžitkov a možných rizík technológie (napr. jadrovej elektrárne) do jedného celkového dobrého alebo zlého pocitu z danej technológie.[76] Napríklad vám najprv poviem, že konkrétny typ jadrového reaktora vytvára menej jadrového odpadu ako iné navrhované reaktory. Ale potom vám

→ http://www.lesswrong.com/lw/gz/policy_debates_should_not_appear_onesided/

76 Melissa L. Finucane et al., „The Affect Heuristic in Judgments of Risks and Benefits,“ *Journal of Behavioral Decision Making* 13, no. 1 (2000): 1–17.

poviem, že tento reaktor je menej stabilný než alternatívne reaktory, s väčším rizikom roztavenia, ak sa dostatočné veľké množstvo vecí pokazí naraz.

Ak má reaktor väčšiu pravdepodobnosť roztavenia, vyzerá to ako „bod proti“ reaktoru, prípadne „bod proti“ tomu, kto argumentuje za postavenie reaktora. A ak ten reaktor vytvára menej odpadu, je to „bod za“ reaktor, prípadne „bod za“ jeho postavenie. Sú teda tieto dva fakty proti sebe? Nie. V skutočnom svete, nie. Tieto dva fakty môžu byť citované rôznymi stranami tej istej diskusie, ale sú logicky nezávislé; fakty nevedia, na čej strane sú. Množstvo odpadu vytvoreného reaktorom závisí od fyzických vlastností daného typu reaktora. Iné fyzické vlastnosti reaktora spôsobujú, že je menej stabilný. Aj keby sa niektoré vlastnosti podieľali na oboch týchto výsledkoch, treba osobitne zvážiť pravdepodobnosť roztavenia a očakávané množstvo odpadu vyprodukované ročne. Sú to dve rôzne fyzikálne otázky, s dvoma rôznymi faktickými odpoveďami.

Štúdie ako horeuvedená však ukazujú, že ľudia majú sklon posudzovať technológie – a mnohé iné problémy – na základe celkového dobrého alebo zlého pocitu. Ak poviete ľuďom, že nejaký typ reaktora produkuje menej odpadu, budú rátať pravdepodobnosť jeho roztavenia ako nižšiu. To znamená, že dostanú *nesprávnu odpoveď* na fyzikálne otázky s konkrétnymi faktickými odpoveďami, pretože zmiešali logicky nezávislé otázky – vnímali fakty ako vojakov bojujúcich na rôznych stranách vojny, a mysleli si, že hocijakého vojaka na jednej strane možno použiť na boj proti hocijakému vojakovi na druhej strane.

Váhy nie sú pre pani Spravodlivosť celkom neprimerané, pokiaľ vyšetruje prísne faktickú otázku viny či neviny. Buď John Smith zabil Johna Doa, alebo nie. Učili nás (E. T. Jaynes), že všetky bayesovské indície sa skladajú z tokov pravdepodobnosti *medzi* hypotézami; že neexistuje nič také ako indícia „podporujúca“ alebo „protirečiaca“ nejakej hypotéze, jedine v tom, zmysle, že iné hypotézy si oproti nej pohoršia alebo prilepšia. Dokiaľ teda pani Spravodlivosť vyšetruje *jednu*, prísne *faktickú* otázku, s iba *dvoma* možnými odpoveďami, váhy sú primeraný nástroj. Ak chce Justitia zvážiť nejakú zložitejšiu tému, mala by sa vzdať svojich váh i svojho meča.

Nie všetky argumenty možno zredukovať na púhe hore alebo dole. Pani Rozumnosť si nosí zápisník, do ktorého si zapisuje všetky fakty, ktoré nie sú na nikoho strane.

*→

—

60. Chyba prisudzovania

> Chyba prisudzovania je sklon vyvodzovať závery o jedinečnej a trvalej povahe osoby na základe správania, ktoré sa dá úplne vysvetliť situáciou, v ktorej nastalo.
>
> --Gilbert a Malone[77]

Máme sklon vidieť príliš priamu súvislosť medzi činmi a osobnosťami druhých ľudí. Keď vidíme, ako niekto iný kope do automatu bez viditeľného dôvodu, predpokladáme, že je to „zurvalec“. Avšak keď my sami kopneme do automatu, je to preto, lebo meškú autobus, ušiel nám vlak, nestíhame prácu, a ešte aj tento prekliaty automat nám zožral peniaze už druhý deň po sebe. Myslíte si, že *v danej situácii by do toho automatu určite kopol každý.*

Svoje vlastné konanie pripisujeme svojej *situácii*, vnímame ho ako dokonale normálnu reakciu na okolnosti. Ale keď niekto iný kopne do automatu, nevidíme za ním vo vzduchu viať jeho minulosť. Vidíme len to kopnutie, *nevieme* o žiadnom dôvode a myslíme si, že to musí byť od prírody zlostný človek – veď predsa útočí bez akejkoľvek provokácie.

Zvážte však pôvodné pravdepodobnosti. Na svete je viac meškajúcich autobusov než mutantov obdarených neprirodzene vysokými hladinami hnevu, ktoré ich nútia občas spontánne nakopnúť automat. Iste, priemerný človek je v skutočnosti mutant. Ak si dobre spomínam, priemerný jednotlivec má 2 až 10

→ http://www.lesswrong.com/lw/h1/the_scales_of_justice_the_notebook_of_rationality/

77 Daniel T. Gilbert and Patrick S. Malone, „The Correspondence Bias,“ *Psychological Bulletin* 117, no. 1 (1995): 21–38, http://www.wjh.harvard.edu/~dtg/Gilbert%20&%20Malone%20(CORRESPONDENCE%20BIAS).pdf.

somaticky vyjadrených mutácií. Ale ľubovoľného *konkrétneho* miesta v DNA sa to týka veľmi nepravdepodobne. Podobne, ľubovoľná stránka niekoho povahy sa pravdepodobne príliš nevzďaľuje od priemeru. Naznačovať opak znamená brať si na plecia bremeno nepravdepodobnosti.

Aj keď ľudí explicitne informujeme o situačných podnetoch, zdá sa, že to nevedia správne odrátať od pozorovaného správania. Keď pokusným osobám povedali, že rečníkom *náhodne pridelili* prečítať prejav za interrupcie alebo proti, pokusné osoby si stále mysleli, že rečník má postoje smerom k náhodne pridelenej téme.[78]

Zdá sa, že je veľmi intuitívne vysvetľovať dážď pomocou vodných duchov; vysvetľovať oheň pomocou ohnivej látky (flogistonu) unikajúcej z horiacej hmoty; vysvetľovať uspávajúci účinok lieku povedaním, že obsahuje „dormitívnu potenciu“. Skutočnosť zvyčajne zahŕňa zložitejšie mechanizmy: cyklus vyparovania a zrážania je základom dažďa; oxidačné horenie je základom ohňa; chemické interakcie s nervovým systémom sú základom uspávadla. Ale mechanizmy nám znejú omnoho zložitejšie než esencie; ťažšie sa na ne myslí, sú menej dostupné. Takže keď niekto kopne do automatu, myslíme si, že má vrodený sklon kopať do automatu.

Okrem prípadu, keď ten „niekto“, kto kopne do automatu, sme my – v tom prípade sa správame dokonale normálne za danej situácie; určite by každý urobil to isté. Veru, preceňujeme sklon druhých reagovať rovnako ako my – „efekt falošného konsenzu“. Študenti pijúci alkohol výrazne preceňujú percento spolužiakov, ktorí pijú; abstinenti ho však výrazne podceňujú. Táto „základná chyba prisudzovania“ označuje náš sklon pripisovať správanie druhých ich povahe, zatiaľ čo u seba naopak.

Aby sme pochopili, prečo ľudia konajú tak, ako konajú, musíme si najprv uvedomiť, že každý vidí sám seba ako normálne sa správajúceho. Nepýtajte sa, s akou zvláštnou mutantskou vlastnosťou priamo zodpovedajúcou ich vonkajšiemu správaniu sa narodili. Radšej sa pýtajte, v akej situácii sa títo ľudia podľa svojho vlastného názoru nachádzajú. Áno, ľudia majú aj vrodené vlastnosti – ale neexistuje *toľko* dedičných vrtochov, aby priamo zodpovedali každému pozorovanému vonkajšiemu správaniu.

Predstavte si, že by som vám dal ovládač s dvoma tlačidlami: červeným a zeleným. Červené tlačidlo zničí celý svet a zelené tlačidlo zabráni stlačeniu červeného tlačidla. Ktoré tlačidlo by ste stlačili? To zelené. Ktokoľvek dá odlišnú odpoveď, pravdepodobne túto otázku príliš komplikuje.→

A predsa sa ma ľudia občas pýtajú, prečo chcem zachrániť svet.→ Akoby som musel mať traumatické detstvo alebo niečo také. Naozaj, vyzerá to ako úplne samozrejmé rozhodnutie... ak vidíte situáciu takýmto pohľadom.

Možno mám netypické názory, ktoré si žiadajú vysvetlenie – prečo verím v takéto veci, keď väčšina ľudí nie? – ale za predpokladu týchto názorov moje *reakcie* snáď nevyžadujú ďalšie vysvetľovanie. Možno som obeťou falošného konsenzu; možno preceňujem koľko ľudí by stlačilo zelené tlačidlo, keby videli situáciu týmto pohľadom. Ale viete, stále by som sa stavil, že sú prinajmenšom vo *výraznej menšine*.

Väčšina ľudí sa zvnútra vníma ako úplne normálnych. Dokonca aj ľudia, ktorých nenávidíte, ľudia, ktorí urobili hrozné veci, nie sú výnimoční mutanti. Žiaľ, na toto mutácie netreba. Ak pochopíte toto, ste pripravení prestať byť prekvapení ľudskými udalosťami.

*→

61. Sú vaši nepriatelia od narodenia zlí?

Vidíme príliš priamu súvislosť medzi konaním druhých a ich vrodenou povahou. Vidíme nezvyčajnú povahu, ktorá presne zodpovedá nezvyčajnému správaniu, namiesto toho, aby sme pátrali

78 Edward E. Jones and Victor A. Harris, „The Attribution of Attitudes,“ [Prisudzovanie vlastností] *Journal of Experimental Social Psychology* 3 (1967): 1–24, http://www.radford.edu/~jaspelme/443/spring-2007/Articles/-Jones_n_Harris_1967.pdf.

→ http://www.yudkowsky.net/singularity/simplified

→ http://intelligence.org/files/AIPosNegFactor.pdf

→ http://www.lesswrong.com/lw/hz/correspondence_bias/

po skutočných alebo imaginárnych situáciách, ktoré by toto správanie mohli vysvetliť. Predpokladáme mutantov.

Keď nám niekto naozaj *ublíži* – spácha skutok, ktorý (právom či neprávom) odsudzujeme – vtedy sa mi zdá, že sa chyba prisudzovania ešte zdvojnásobí. Máme zrejme *veľmi* silný sklon obviňovať za zlé skutky nepriateľovu mutantnú, zlú povahu. Nie z morálneho hľadiska, ale ako technickú otázku pôvodnej pravdepodobnosti, by sme sa mali opýtať, čo si nepriateľ môže myslieť o svojej situácii, čo by mohlo zmenšiť zdanlivú čudesnosť jeho správania. To by nám umožnilo prísť s hypotézou menej výnimočnej povahy, a tak by sme niesli menšie bremeno nepravdepodobnosti.

11. septembra 2001 deväťnásť moslimských mužov unieslo štyri lietadlá so samovražedným zámerom uškodiť Spojeným Štátom Americkým. Čo si myslíte, prečo to mohli urobiť? Pretože videli USA ako maják svetovej slobody, ale narodili sa s mutantnou povahou, ktorá spôsobuje, že nenávidia slobodu?

Realisticky, väčšina ľudí si neskladá svoj životný príbeh so sebou samým v roli záporného hrdinu. Každý je kladným hrdinom vo svojom vlastnom príbehu. V nepriateľovom príbehu, z pohľadu nepriateľa, *nebude nepriateľ vykreslený ako ten zlý.* Ak sa snažíte vyskladať motiváciu, podľa ktorej *má* nepriateľ vyzerať ako ten zlý, budete sa úplne mýliť ohľadom toho, čo sa v nepriateľovej hlave naozaj odohráva.

Lenže politika zabíja myslenie. Debata je vojna, argumenty sú vojaci. Akonáhle viete, na ktorej strane ste, musíte podporovať všetky argumenty tejto strany, a napádať všetky argumenty, ktoré vyzerajú v prospech nepriateľskej strany; inak je to, ako keby ste vlastných vojakov bodali do chrbta.

Keby nepriateľ mal zlú povahu, bol by to argument v prospech vašej strany. A *každý* argument v prospech vašej strany musíte podporiť, bez ohľadu na jeho hlúposť – inak prestávate tlačiť v nejakej časti frontovej línie. Každý sa snaží prekonať svojho suseda vlasteneckými prehláseniami a nikto sa neopovažuje protirečiť. Čoskoro má nepriateľ rohy, netopierie krídla, ohnivý dych a pazúre, z ktorých odkvapkáva jedovatá kyselina. Ak sa pokúsite niečo z tohto poprieť na základe púhych faktov, argumentujete za nepriateľovu stranu; ste zradca. Veľmi málo ľudí dokáže pochopiť, že nebránite nepriateľa, iba bránite pravdu.

Keby bolo treba mutantov na to, aby robili príšerné veci, história ľudského druhu by vyzerala veľmi odlišne. Mutantov by bolo málo.

Alebo je to možno strach, že porozumenie povedie k odpusteniu. Je ľahšie strieľať do zlých mutantov. O čo inšpirujúcejšie znie bojový pokrik: „Skapte, mizerní lotri!“ než „Skapte, ľudia, ktorí mohli byť takí istí ako ja, ale vyrástli v odlišnom prostredí!“ Zo zabíjania ľudí, ktorí *neboli* čistou temnotou, by ste mohli mať pocit viny.

Toto mi pripadá ako hlboko zakorenená túžba po jednostrannej debate o pravidlách, v ktorej to najlepšie rozhodnutie nemá *žiadne* nevýhody. Ak armáda prekračuje hranice, alebo k vám ide šialenec s nožom, alternatívy sú: (a) brániť sa, (b) ľahnúť si a zomrieť. Ak sa budete brániť, možno budete musieť zabiť. Ak zabijete niekoho, kto mohol, v inom svete, byť vaším priateľom, je to tragédia. Áno, *je* to tragédia. Druhá možnosť, ľahnúť si a zomrieť, je tiež tragédia. Prečo by musela existovať aj netragická možnosť? Kto hovorí, že najlepšie možné rozhodnutie nesmie mať žiadne nevýhody? Ak už niekto musí zomrieť, rovnako dobre to môže byť iniciátor útoku, aby sme tým odradili budúce násilie a tým minimalizovali celkové množstvo smrti.

Ak má nepriateľ priemernú povahu a koná podľa takého názoru na svoju situáciu, pri ktorom je násilie typickou ľudskou reakciou, to ešte neznamená, že jeho názory sú fakticky správne. Neznamená to, že sú oprávnené. Znamená to, že budete musieť zastreliť niekoho, kto je vo svojom vlastnom príbehu kladným hrdinom a v ich príbehu kladný hrdina zomrie na strane 80. To je tragédia, ale je to lepšie než alternatívna tragédia. Je to voľba, aké robí každý policajt, každý deň, aby sa naše milé malé svety nerozpustili v chaose.

Keď presne odhadujete nepriateľovu psychológiu – keď viete, čo naozaj je v nepriateľovej hlave – toto poznanie vám nedá pocit, že zasadzujete nádherný úder súperovi. Nedá vám hrejivý pocit spravodlivého rozhorčenia. Nebudete mať vďaka nemu dobrý pocit zo seba samého. Ak vám váš odhad prináša neznesiteľný smútok, možno vidíte svet taký, aký naozaj je. Zriedkavejšie sa môže stať, že vám

presný odhad privolá zimomriavky vážnej hrôzy, ako keď jednáte so skutočnými psychopatmi alebo s neurologicky nepoškodenými ľuďmi s názormi, ktoré naprosto zničili ich príčetnosť (scientológovia alebo Jesus Camp).

Povedzme to teda priamo a nahlas – únoscovia z 11. 9. neboli zlí mutanti. Necítili nenávisť k slobode. Aj oni boli kladnými hrdinami vo svojich vlastných príbehoch a zomreli za to, čo považovali za správne – pravdu, spravodlivosť, islamský spôsob života. To, že sa tak sami vnímali, neznamená, že mali pravdu. To, že sa tak vnímali, neznamená, že musíme súhlasiť, že mali právo urobiť, čo urobili. To, že sa tak vnímali, neznamená, že pasažieri United Flight 93 mali ustúpiť a nechať im voľnú ruku. Znamená to, že v inom svete, keby vyrástli v inom prostredí, títo únoscovia mohli byť napríklad policajtmi. A to je naozaj tragédia. Vitajte na Zemi.

*→

62. Obrátená hlúposť nie je inteligencia

> „...potom naši ľudia v tejto časovej línii prišli s nápravnou akciou. Tu.“
>
> Utrel obrazovku a začal zadávať kombinácie. Objavovala sa strana za stranou s údajmi o ľuďoch, ktorí tvrdili, že videli tajomné disky a každá správa bola fantastickejšia než predchádzajúca.
>
> „Štandardná zahladzovacia technika,“ uškrnul sa Verkan Vall. „Počul som iba pár zmienok o ‚lietajúcich tanieroch‘ a všetko to bolo žartom. V kultúre tohto typu dokážete vždy urobiť pravdivý príbeh nedôveryhodným, ak vedľa neho postavíte desať ďalších, jasne nepravdivých.“
>
> --H. Beam Piper, Policajná operácia[79]

Piper má v niečom pravdu. Ja osobne neverím na žiadnych zle ukrytých mimozemšťanov v našich končinách. Ale moja neviera nemá nič spoločné so zahanbujúcou nerozumnosťou kultov lietajúcich tanierov – prinajmenšom dúfam, že nie.

Vy a ja veríme, že kulty lietajúcich tanierov vznikli v úplnej neprítomnosti lietajúcich tanierov. Kulty dokážu vďaka ľudskej hlúposti vzniknúť okolo takmer hocijakej myšlienky. Táto hlúposť funguje *nezávisle* na mimozemských zásahoch: Predpokladali by sme, že kulty lietajúcich tanierov vzniknú bez ohľadu na to, či lietajúce taniere existujú alebo nie. Dokonca aj keby existovali zle ukrytí mimozemšťania, nebolo by o nič *menej* pravdepodobné, že vzniknú kulty lietajúcich tanierov. P(kulty|mimozemšťania) nie je menšie ako P(kulty|~mimozemšťania), pokiaľ nepredpokladáte, že zle ukrytí mimozemšťania by úmyselne potláčali kulty lietajúcich tanierov. Podľa bayesovskej definície indície, pozorovanie „existujú kulty lietajúcich tanierov“ nie je indícia *proti* existencii lietajúcich tanierov. Nie je to veľmi indícia jedným ani druhým smerom.

Toto je použitie všeobecného princípu, podľa ktorého, ako hovorí Robert Pirsig: „Najväčší hlupák na svete môže povedať, že Slnko svieti, ale ono kvôli tomu nezhasne.“[80]

Keby ste vedeli, že sa niekto mýli v 99,99 % prípadov otázok typu áno alebo nie, mohli by ste získať presnosť 99,99 % jednoducho obrátením jeho odpovedí. On by musel vynaložiť všetku prácu potrebnú na získanie dobrých indícií previazaných so skutočnosťou a systematicky tieto indície spracovať, aby dokázal takto spoľahlivo *antikorelovať*. Musel by byť superinteligentný, aby mohol byť taký hlúpy.

Auto s pokazeným motorom nebude cúvať rýchlosťou 300 km/h, ani keď je ten motor *naozaj, naozaj pokazený.*

→ http://lesswrong.com/lw/i0/are_your_enemies_innately_evil/

79 Henry Beam Piper, „Police Operation,“ Astounding Science Fiction (July 1948).

80 Robert M. Pirsig, *Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values*, 1st ed. (New York: Morrow, 1974).

Ak hlúposť nedokáže spoľahlivo antikorelovať s pravdou, o čo menej by malo s pravdou antikorelovať ľudské zlo? Opakom efektu svätožiary je efekt rohov. Všetky vnímané záporné vlastnosti korelujú. Ak je Stalin zlý, všetko, čo povie, musí byť nepravdivé. Nechceli by ste predsa súhlasiť *so Stalinom*, alebo áno?

Stalin si myslel, že 2 + 2 = 4. Ak však budete obhajovať akýkoľvek Stalinov výrok, vrátane „2 + 2 = 4“, ľudia budú vidieť iba to, že „súhlasíte so Stalinom“; čiže musíte byť na jeho strane.

Dôsledky tohto princípu:

- Ak chcete poctivo argumentovať proti nejakej myšlienke, mali by ste argumentovať proti najlepším argumentom jej najsilnejších zástancov. Argumentovanie proti slabším zástancom nedokazuje *nič*, pretože aj tá najsilnejšia myšlienka bude priťahovať slabých zástancov. Ak chcete argumentovať proti transhumanizmu alebo explózii inteligencie, musíte priamo čeliť argumentom Nicka Bostroma alebo Eliezera Yudkowskeho po roku 2003. Najmenej pohodlná cesta je tá jediná správna.
- Ukazovať na smutných trápnych bláznov dohnaných k šialenstvu ich chápaním nejakej myšlienky nie je indícia proti tejto myšlienka. Mnoho priaznivcov New Age sa stalo ešte bláznivejšími vďaka svojmu osobnému výkladu kvantovej mechaniky.
- Niekto raz povedal: „Nie všetci konzervatívci sú hlúpi, ale väčšina hlúpych ľudí sú konzervatívci.“ Ak sa nedokážete dostať do stavu mysle, keď vám tento výrok, či už je pravdivý alebo nie, pripadá *úplne nepodstatný* ako kritika konzervativizmu, potom nie ste pripravení rozumne rozmýšľať o politike.
- Argument ad hominem nie je správny.
- Musíte byť schopní argumentovať proti genocíde bez slov: „Hitler chcel vyvraždiť židov.“ Keby Hitler *nebol chcel* genocídu, znamenalo by to, že genocída je okej?
- Hansonovskými slovami: Vaša inštinktívna ochota veriť nejakej myšlienke sa bude meniť podľa vašej ochoty *spájať sa* s ľuďmi, ktorí sú známi ako jej zástancovia – čo nesúvisí s jej skutočnou *pravdivosťou*. Niektorí ľudia môžu byť neochotní pripustiť, že Boh neexistuje, nie preto, že by existovali dôkazy, že Boh *existuje*, ale skôr preto, lebo sa nechcú pridávať k Richardovi Dawkinsovi ani tým prekliatym „provokujúcim“ ateistom, ktorí chodia a nahlas hovoria: „Boh neexistuje.“
- Ak váš terajší počítač prestane fungovať, nemôžete dôjsť k záveru, že všetko na vašom terajšom systéme je pokazené a že potrebujete nový systém bez procesora AMD, bez videokarty ATI, bez pevného disku Maxtor a bez ventilátora na krabici – aj keď váš terajší počítač všetko toto má a nefunguje. Možno len potrebujete nový kábel od napájania.
- Ak sa sto vynálezcom nepodarí[→] postaviť lietajúci stroj pomocou kovu a dreva a plátna, to neznamená, že v skutočnosti potrebujete lietajúci stroj z mäsa a kostí. Ak sa tisíc projektom nepodarí postaviť umelú inteligenciu pomocou elektronických obvodov, to nedokazuje, že zdrojom problému je elektrina. Dokiaľ nepochopíte problém, naivné obrátenie s najväčšou pravdepodobnosťou k riešeniu nepovedie.

*→

63. Argument zatieňuje autoritu

Scenár 1: Barry je známy geológ. Charles je 14-ročný mladistvý delinkvent s dlhým záznamom trestov a občasnými psychotickými záchvatmi. Barry bez vysvetlenia povie Arthurovi nejaké kontraintuitívne tvrdenie o kameňoch a Arthur tomu prisúdi pravdepodobnosť 90 %. Potom Charles povie rovnako kontraintuitívne tvrdenie o kameňoch a Arthur tomu prisúdi pravdepodobnosť 10 %. Je jasné, že Arthur pri rozhodovaní, či má danému tvrdeniu veriť, zohľadňuje *autoritu* hovoriaceho.

→ http://lesswrong.com/lw/vs/selling_nonapples/

→ http://lesswrong.com/lw/lw/reversed_stupidity_is_not_intelligence/

Scenár 2: David povie kontraintuitívne tvrdenie o fyzike a dá Arthurovi podrobné vysvetlenie argumentov, vrátane odkazov na odbornú literatúru. Ernie povie rovnako kontraintuitívne tvrdenie, ale dá nepresvedčivé vysvetlenie obsahujúce niekoľko nepodložených skokov. Obaja tvrdia, že je to najlepšie vysvetlenie, aké dokážu poskytnúť (hocikomu, nielen Arthurovi). Po vypočutí si vysvetlení, Arthur prisúdi Davidovmu tvrdeniu pravdepodobnosť 90 % a Ernieho tvrdeniu pravdepodobnosť 10 %.

Môže sa zdať, že tieto dva scenáre sú zhruba súmerné: oba obsahujú zváženie užitočnej indície, buď silnej a slabej autority alebo silného a slabého argumentu.

Predstavme si však, že Arthur požiada Barryho a Charlesa o plné technické vysvetlenie vrátane odkazov na odbornú literatúru; a že Barry a Charles predložia rovnako dobré vysvetlenia a Arthur si pozrie odkazy a všetky sedia. Potom Arthur požiada Davida a Ernieho o potvrdenie ich odbornosti a ukáže sa, že David a Ernie majú zhruba rovnakú odbornosť – možno sú obaja šašovia, možno sú obaja fyzici.

Ak predpokladáme, že Arthur má dostatočné vedomosti, aby pochopil všetky technické argumenty – lebo inak je to len dobre pôsobiaci šum – zdá sa, že by Arthur mal dôjsť k záveru, že David je omnoho dôveryhodnejší než Ernie, zatiaľ čo Barry má v lepšom prípade iba maličkú výhodu oproti Charlesovi.

Veru, ak sú technické argumenty dosť dobré, Barryho výhoda voči Charlesovi možno nestojí za reč. Dobrý technický argument je taký, ktorý *odstraňuje* závislosť na osobnej autorite hovoriaceho.

Podobne, ak naozaj veríme, že Ernie dal najlepší argument, aký *mohol* dať, čo zahŕňa všetky jeho inferenčné kroky a všetku podporu, ktorú zohľadnil – vrátane citátov všetkých autorít, ktoré samotný Ernie počul – potom môžeme viacmenej ignorovať informáciu o Ernieho odbornosti. Ernie môže byť fyzik alebo šašo, na tom nezáleží. (Opäť, predpokladáme, že máme dostatočné technické zručnosti na spracovanie argumentu. V opačnom prípade Ernie jednoducho vyslovuje záhadné slabiky a či týmto slabikám „veríme“, závisí do veľkej miery od jeho autority.)

Zdá sa teda, že medzi autoritou a argumentom je nesúmernosť. Ak poznáme autoritu, stále nás zaujímajú argumenty; ak však úplne poznáme argumenty, autorita nám povie už iba máličko.

Zrejme (povie začiatočník) sú autorita a dôkaz dva zásadne odlišné druhy indícií a tento rozdiel nie je zohľadnený v nudne zjednodušených metódach bayesovskej teórie pravdepodobnosti. Pretože hoci je sila indícií v oboch prípadoch rovnaká – 90 % verzus 10 % – nesprávajú sa podobne, keď ich skombinujeme. Ach, ako to len vysvetlíme?

Tu je polovica technickej ukážky, ako tento rozdiel reprezentovať pomocou teórie pravdepodobnosti. (Zvyšok môžete prijať na základe mojej osobnej autority alebo si dohľadať odbornú literatúru.)

Ak P(H|E1) = 90 % a P(H|E2) = 9 %, aká je pravdepodobnosť P(H|E1,E2)? Ak na základe informácie, že E1 je pravda, priradíme H pravdepodobnosť 90 % a na základe informácie, že E2 je pravda, priradíme H pravdepodobnosť 9 %, akú pravdepodobnosť by sme mali priradiť H, ak sa dozvieme, že platí aj E1 aj E2? Toto sa podľa teórie pravdepodobnosti jednoducho nedá vypočítať iba na základe uvedených informácií. Nie, tá chýbajúca informácia nie je apriórna pravdepodobnosť H. E1 a E2 nemusia byť navzájom nezávislé.

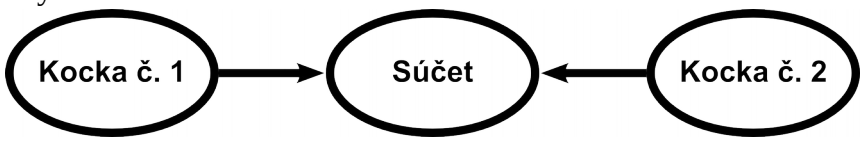
Predpokladajme, že H je „môj chodník sa šmýka“, E1 je „môj zavlažovač beží“ a E2 je „je noc“. Chodník sa začne šmýkať 1 minútu po spustení zavlažovača a prestane hneď ako zavlažovač skončí, a zavlažovač ide 10 minút. Vieme teda, že ak zavlažovač beží, chodník sa šmýka s pravdepodobnosťou 90 %. Zavlažovač je zapnutý 10 % nočného času, ak teda vieme, že je noc, pravdepodobnosť, že chodník sa šmýka, je 9 %. Ak vieme, že je noc a že je zavlažovač zapnutý – teda ak vieme oba tieto fakty – pravdepodobnosť, že chodník sa šmýka, je 90 %.

Môžeme si to znázorniť nasledujúcim grafickým modelom:

Či je alebo nie je noc *spôsobuje* zapnutie a vypnutie zavlažovača, a či je zavlažovač zapnutý *spôsobuje* šmýkanie alebo nešmýkanie chodníka.

Smer šípok je dôležitý. Keby sme mali:

Znamenalo by to, že keby som *nevedel* nič o zavlažovači, pravdepodobnosť noci a šmykľavosti by boli navzájom nezávislé. Napríklad, predstavme si, že hodím kockou číslo jedna a kockou číslo dva a získané čísla sčítam, aby som dostal súčet:



Ak mi nepoviete súčet dvoch čísel, a ak mi poviete, že na prvej kocke padlo 6, to mi nič nehovorí o výsledku na druhej kocke, zatiaľ. Ak mi však zároveň poviete, že súčet je 7, viem, že na druhej kocke padlo 1.

Zisťovanie, či sú rôzne kúsky informácie navzájom závislé alebo nezávislé, za rôznych všeobecných podmienok, sa ukazuje ako samostatná technická téma. Vhodné knihy sú od Judeu Pearla Pravdepodobnostné uvažovanie v inteligentných systémoch: Siete dôveryhodného odvodzovania[81] a Kauzalita[82]. (Ak máte čas iba na jednu knihu, prečítajte si tú prvú.)

Ak viete čítať kauzálne grafy, potom sa pozriete na graf hádzania kociek a hneď vidíte:

p(kocka1, kocka2) = p(kocka1) × p(kocka2)

p(kocka1, kocka2 | súčet) ≠ p(kocka1 | súčet) × p(kocka2 | súčet)

Ak sa pozriete na správny diagram chodníka, uvidíte fakty ako:

p(šmýkanie | noc) ≠ p(šmýkanie)

p(šmýkanie | zavlažovač) ≠ p(šmýkanie)

p(šmýkanie | noc, zavlažovač) = p(šmýkanie | zavlažovač)

To znamená, že pravdepodobnosť, že sa chodník šmýka, ak vieme o zavlažovači a noci, je rovnaká ako pravdepodobnosť, ktorú by sme priradili, keby sme vedeli iba o zavlažovači. Vedomosť o zavlažovači urobila vedomosť o noci nepodstatnou na usudzovanie o šmykľavosti.

Toto sa nazýva *zatienenie* a kritérium, ktoré nám dovoľuje vyčítať takéto podmienené nezávislosti z kauzálnych grafov sa nazýva *D-separácia*.

V prípade argumentu a autority vyzerá kauzálny diagram takto:

Ak je niečo pravda, zvyknú v prospech toho existovať argumenty, odborníci preto vidia tieto indície a menia svoje názory. (Teoreticky!)

Ak vidíme, že si odborník niečo myslí, môžeme spätne usudzovať existenciu nejakej indície (aj keď nevieme, čo presne je táto indícia) a z existencie tejto hypotetickej indície môžeme spätne usudzovať pravdivosť výroku.

Ak však poznáme hodnotu uzla „Argument“, uzol „Pravda“ sa tým D-separuje od uzla „Názor odborníka“ zablokovaním všetkých ciest medzi nimi, podľa určitého technického kritéria pre „zablokovanie cesty“, ktoré v tomto prípade vyzerá celkom jasne. Takže aj bez kontroly presnej distribúcie pravdepodobnosti môžeme z grafu vyčítať, že:

p(pravda | argument, odborník) = p(pravda | argument)

Toto nie je v rozpore s bežnou teóriou pravdepodobnosti. Je to iba kompaktnejší spôsob, ako vyjadriť určité pravdepodobnostné fakty. Dokázali by ste vyčítať tie isté rovnice a nerovnice

81 Pearl, *Probabilistic Reasoning in Intelligent Systems*.

82 Judea Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. (New York: Cambridge University Press, 2009).

z neozdobenej pravdepodobnostnej distribúcie – ale bolo by ťažšie to priamo uvidieť. Autorita a argument nevyžadujú dva rôzne druhy pravdepodobnosti, rovnako ako zavlažovače nemajú ontologicky odlišnú podstatu než slnečné svetlo.

V praxi sa závislosti na autorite nikdy *celkom* nezbavíte. Dobré autority budú s väčšou pravdepodobnosťou poznať existujúce indície proti niečomu a to by ste mali zohľadniť; menšia autorita ich bude poznať s menšou pravdepodobnosťou, čo robí jej argumenty menej spoľahlivé. Toto nie je faktor, ktorý môžete odstrániť púhym vypočutím si indícií, ktoré *zohľadnili*.

Je tiež veľmi ťažké redukovať argumenty na *číru* matematiku; navyše, posudzovanie sily inferenčného kroku môže navyše závisieť na intuíciách, ktoré nedokážete duplikovať bez rovnakých tridsiatich rokov skúsenosti.

Je nezmazateľne legitímne priradiť *trochu* vyššiu pravdepodobnosť tomu, čo vám o bayesovskej pravdepodobnosti hovorí E. T. Jaynes, než priraďujete celkom rovnakému výroku, ktorý povie Eliezer Yudkowsky. Päťdesiat rokov skúsenosti navyše by sa nemali počítať ako doslova *nulový* vplyv.

Ale táto o čosi silnejšia autorita je iba *ceteris paribus* a dá sa ľahko prevážiť silnejšími argumentmi. Našiel som v jednej z Jaynesových kníh menší preklep – pretože algebra prebíja autoritu.

* →

64. Pritisnite si otázku

V umení rozumnosti existuje zručnosť *blízkosti k téme* – snažiť sa sledovať indície, ktoré sú k pôvodnej otázke čo najbližšie, takže zatieňujú čo najviac iných argumentov.

Bratia Wrightovci povedia: „Moje lietadlo poletí.“ Ak sa pozriete na ich autoritu (opravári bicyklov, ktorí sú zhodou okolností vynikajúci amatérski fyzici) a porovnáte ju s autoritou povedzme Lorda Kelvina, zistíte, že Lord Kelvin má väčšiu autoritu.

Ak si vypýtate na nahliadnutie výpočty bratov Wrightovcov a budete im rozumieť, a ak si vypýtate výpočty Lorda Kelvina (pravdepodobne nemá žiadne okrem vlastnej nedôverčivosti), potom bude autorita omnoho menej dôležitá.

Ak naozaj *uvidíte, ako to lietadlo letí*, samotné výpočty sa stanú na mnohé účely nepodstatnými a Kelvinova autorita nebude stáť za zmienku.

Čím *priamejšie* sa vaše argumenty týkajú nejakej otázky, bez sprostredkujúcich vplyvov – čím bližšie sú pozorované uzly k uzlu otázky vo Veľkej Sieti Kauzality – tým mocnejšia je indícia. V týchto kauzálnych grafoch platí veta, že nikdy nemôžete dostať *viac* informácií zo vzdialenejších uzlov než z bližších uzlov, ktoré tie vzdialenejšie zatieňujú.

Jerry Cleaver povedal: „To, čo ťa zloží, nie je zabudnutie použiť nejakú zložitú neprehľadnú techniku na vysokej úrovni. Je to nevšímanie si základov. Nesledovanie lopty pohľadom.“[83]

Rovnako ako je lepšie argumentovať fyzikou než diplomom, je lepšie argumentovať fyzikou než rozumnosťou. Kto bol rozumnejší, bratia Wrightovci alebo Lord Kelvin? Ak môžeme skontrolovať ich rovnice, môže nám to byť jedno! Cnosť racionalistu nedokáže *priamo* zdvihnúť lietadlo do vzduchu.

Ak zabudnete na tento princíp, bude vám učenie sa o ďalších skresleniach škodiť, pretože vás bude rozptyľovať od priamejších argumentov. Je príliš ľahké argumentovať, že niekto vykazuje Skreslenie číslo 182 vo vašom repertoári úplne všeobecných obvinení, ale bez bližšej indície nemožno *uzavrieť* faktickú tému. Ak existujú skreslené dôvody hovoriť, že slnko svieti, to ešte nespôsobuje, že zhasne.

Rovnako ako nemôžete vždy urobiť experiment dnes, nemôžete vždy skontrolovať rovnice dnes. Niekedy dostatočne nepoznáte pozadie, niekedy sa informácia nedá odovzdať, niekedy jednoducho nie je čas. Je žiaľ veľa prípadov, keď sa oplatí hodnotiť rozumnosť hovoriaceho. Vždy by ste to však mali robiť s pocitom prázdnoty v srdci, s vedomím, že niečo tu chýba.

→ http://lesswrong.com/lw/lx/argument_screens_off_authority/

83 Jerry Cleaver, *Immediate Fiction: A Complete Writing Course* (Macmillan, 2004).

Vždy keď môžete, tancuje tak blízko pôvodnej otázky, ako sa dá – pritlačte sa k nej – dostaňte sa blízko a *pritisnite si ju!*

*→

65. Rozumnosť a anglický jazyk

V reakcii na moje písanie o signáloch na potlesk mi niekto povedal, že mu moje písanie pripomenulo Politika a anglický jazyk[84][84] od Georgea Orwella. Je mi to cťou. Najmä preto, lebo som už mal vymyslenú dnešnú tému.

Ak *naozaj* chcete umelcov pohľad na rozumnosť, potom si prečítajte Orwella; je to povinné čítanie pre racionalistov aj pre spisovateľov. Orwell nebol vedec, ale spisovateľ; jeho nástrojmi neboli čísla, ale slová; jeho protivníkom nebola Príroda, ale ľudské zlo. Ak chcete uväzniť ľudí na celé roky bez súdneho procesu, musíte si vymyslieť iný spôsob, ako to povedať, než: „Idem uväzniť pána Jenningsa na celé roky bez súdneho procesu.“ Musíte zahmliť poslucháčovo myslenie, zabrániť tomu, aby jasný obraz pobúril jeho svedomie. Poviete: „Nespoľahlivé živly boli podrobené alternatívnemu súdnemu procesu.“

Orwell bol pobúreným protivníkom totalitárstva a zahmleného myslenia, do ktorého sa zlo zahaľuje – a tak Orwellove písania o jazyku skončili ako klasické racionalistické dokumenty na úrovni Feynmana, Sagana alebo Dawkinsa.

„Spisovateľom sa hovorí, aby sa vyhli používaniu pasívneho tvaru.“ Racionalista, ktorý je podkutý *výhradne* vo vede, si nemusí všimnúť chybu v predchádzajúcej vete; ale každý, kto trochu písal, by ju mal uvidieť hneď. Napísal som tú vetu v pasívnom tvare; nepovedal som vám, *kto* hovorí autorom, aby sa vyhli pasívnemu tvaru. Pasívny tvar odstraňuje konajúceho, zanecháva iba konané. „Nespoľahlivé živly boli podrobené alternatívnemu súdnemu procesu“ – podrobené *kým*? Ako sa *robí* taký „alternatívny súdny proces“? S dostatkom statických menných pomenovaní sa dokážete *vyhnúť* čomukoľvek nepríjemnému.

Články v časopisoch sú často písané v pasívnom tvare. (Prepáčte; *niektorí vedci* píšu svoje články do časopisov v pasívnom tvare. Tie články sa predsa nepíšu sami.) Znie autoritatívnejšie, keď poviete „pokusným osobám bol podávaný Progenitorivox“, než „dal som každému vysokoškolákovi fľašku s 20 Progenitorivoxmi a povedal som mu, nech si každý večer dá jeden, dokiaľ sa neminú.“ Ak odstránite vedca z popisu, zostane vám iba to najdôležitejšie, údaje. Lenže v skutočnosti tam ten vedec *je* a pokusné osoby *sú* vysokoškoláci a Progenitorivox nebol „podávaný“ ale odovzdaný s pokynmi. Pasívny tvar zastiera skutočnosť.

Súdiac podľa komentárov, ktoré dostávam na Overcoming Bias, niekto bude namietať, že používanie pasívneho tvaru v článku časopisu je sotva hriechom – napokon, keď sa nad tým *zamyslíte*, uvedomíte si, že vedec je tam. Nevyzerá to ako logická chyba. A to je dôvod, prečo racionalisti potrebujú čítať Orwella, nielen Feynmana či Jaynesa.

Literatúra faktu sprostredkováva *poznanie*, fikcia sprostredkováva *zážitok*. Medicína dokáže extrapolovať, čo by sa stalo nechránenému človeku vo vzduchoprázdne. Fikcia dokáže, aby ste si to prežili.

Niektorí racionalisti sa pokúsia analyzovať zavádzajúcu vetu, pokúsia sa vidieť, či by tam *mohlo byť* niečo zmysluplné, pokúsia sa *zostrojiť* logickú interpretáciu. Budú zhovievaví, v prípade pochybnosti sa priklonia na stranu autora. Na druhej strane, spisovatelia sa cvičia v tom, aby k sebe *neboli* zhovievaví. Čokoľvek si obecenstvo *myslí*, že ste povedali, *to* ste povedali, či už ste to chceli povedať alebo nie; nemôžete sa hádať s obecenstvom bez ohľadu na to, aké chytré sú vaše zdôvodnenia.

Spisovateľ vie, že čitateľ sa *nezastaví*, aby chvíľu porozmýšľal. Zážitok z fikcie je nepretržitý prúd prvých dojmov. Spisovateľ-racionalista venuje pozornosť *zážitkom*, ktoré slová vytvárajú. Ak vyhodnocujete verejnú rozumnosť nejakej vety, ak zámerne analyzujete slová, preformulovávate

→ http://lesswrong.com/lw/ly/hug_the_query/

84 George Orwell, „Politics and the English Language,“ *Horizon* (April 1946).

tvrdenia, skúšate rôzne významy, hľadáte zrnká pravdy, potom strácate zo zreteľa prvý dojem – čo obecenstvo *vidí* alebo skôr, čo *cíti*.

Románopisec by si všimol krikľavú nesprávnosť vety: „Pokusným osobám bol podávaný Progenitorivox“. Kde je tam niečo živé, čo môže čitateľ prežívať? Táto veta vytvára vzdialený pocit autoritatívnosti a to je *všetko – jediným* zážitkom je pocit, že vám niekto povedal niečo spoľahlivé. Románopisec by videl podstatné mená ako príliš abstraktné na to, aby ukázali, čo sa naozaj stalo – postdoktorand s fľaškou v ruke, snažiaci sa tváriť prísne; študent počúvajúci s nervóznym úškrnom.

Nechcem tým povedať, že by sa články do časopisov mali písať ako romány, ale že racionalista by si mal začať uvedomovať, aké *zážitky* ktoré slová vytvárajú. Racionalista musí chápať myseľ a ako ju ovládať. To zahŕňa aj prúd vedomia, tú časť vás, ktorá sa rozvíja v jazyku. Racionalista si musí začať uvedomovať skutočný, skúsenostný dopad viet, nielen ich púhu výrokovú sémantiku.

Alebo, stručnejšie povedané: *Význam neospravedlňuje dopad!*

Nezaujíma ma, aká racionálne interpretácie dokážete *zostrojiť* zo signálov na potlesk ako: „UI by mala byť vyvinutá demokratickým procesom“. To neospravedlňuje iracionálny dopad tejto vety, signalizovanie obecenstvu, aby zatlieskalo, nehovoriac o jej pochybnej nejasnosti.

Tu je Orwell, karhajúci *dopad* fráz, ich účinok na zážitok myslenia:

> Keď človek sleduje nejakého unaveného kecala na pódiu, ako mechanicky opakuje známe frázy – *zverstvá*, *zločiny*, *železná čižma*, *krvavá tyrania*, *slobodní ľudia celého sveta*, *stojme bok po boku* – často má čudný pocit, že nesleduje živú ľudskú bytosť, ale akúsi bábku... Rečník, ktorý používa tento druh frazeológie, sa do istej miery sám premenil na stroj. Z jeho hrtana vychádzajú príslušné zvuky, ale jeho mozog sa na tom nezúčastňuje, nie ako keby tie slová vyberal sám...
>
> Predovšetkým treba nechať význam, nech si vyberie slovo, a nie naopak. Najhoršia vec, ktorú v próze môžete urobiť so slovami, je poddať sa im. Keď myslíte na konkrétny predmet, myslíte bez slov a potom, keď chcete popísať tú vec, ktorú ste si predstavovali, asi lovíte, dokiaľ nenájdete tie správne slová, ktoré vyzerajú primerane. Keď myslíte na niečo abstraktné, máte väčší sklon začať od slov a pokiaľ nevynaložíte vedomé úsilie, aby ste tomu zabránili, existujúci dialekt vtrhne a vykoná svoje dielo za vás, na úkor zahmlenia alebo priam zmenenia toho, čo ste mysleli. Je pravdepodobne lepšie odkladať používanie slov tak dlho, ako sa len dá, a čo najviac si vyjasniť svoj význam pomocou obrazov a pocitov.

Posledný odsek akoby napísal Charles Sanders Peirce. K Ceste vedie viac než jedna cesta.

* →

66. Ľudské zlo a zahmlené myslenie

George Orwell videl pád civilizovaného sveta do totalitárstva, premenu alebo úpadok jednej krajiny za druhou; čižma dupajúca na ľudskú tvár, naveky a pamätaj, že naveky. Vy ste sa narodili príliš neskoro na to, aby ste si pamätali časy, keď rast totalitárstva vyzeral nezastaviteľný, keď jedna krajina za druhou padala do rúk tajnej polície, hromové klopanie o polnoci, zatiaľ čo profesori na univerzitách slobodného sveta oslavovali čistky v Sovietskom Zväze ako pokrok. Pripadá vám to rovnako cudzie ako fikcia; je ťažké brať to vážne. Pretože vo vašej vetve času berlínsky múr padol. A ak Orwellovo meno nie je vytesané do jedného z tých kameňov, malo by byť.

Orwell videl osud ľudského druhu a vynaložil kŕčovité úsilie, aby ho vytrhol z jeho cesty. Orwellovou zbraňou bolo jasné písanie. Orwell vedel, že zmätený jazyk je zmätené myslenie; vedel, že ľudské zlo a zmätené myslenie sa prepletajú ako konjugované vlákna DNA:[85]

→ http://www.lesswrong.com/lw/jc/rationality_and_the_english_language/

85 Orwell, „Politics and the English Language.“

> V našej dobe je politická reč a písanie prevažne obhajovaním neobhájiteľného. Veci ako pokračovanie britskej nadvlády v Indii, ruské čistky a deportácie, zhodenie atómových bômb na Japonsko sa síce dajú obhajovať, ale iba argumentmi, ktoré väčšine ľudí pripadajú príliš brutálne a ktoré nie sú v súlade s vyhlasovanými cieľmi politických strán. Preto sa politický jazyk musí skladať prevažne z eufemizmov, podsúvania odpovedí a čírej hmlistej nejasnosti. Bezbranné dediny sú bombardované zo vzduchu, ich obyvatelia sú vyhnaní do polí, dobytok postrieľaný samopalmi, chyže podpálené zápalnými strelami: toto sa nazýva *pacifikovanie*...

Orwell bol priamy ohľadom cieľa svojej jasnosti:

> Ak zjednodušíte svoju angličtinu, oslobodíte sa od najhorších bláznovstiev pravovernosti. Nemôžete hovoriť žiadnym z jej nevyhnutných nárečí a keď urobíte hlúpu poznámku, jej hlúposť bude zrejmá, dokonca aj vám samotným.

Urobiť našu hlúposť zrejmou, dokonca aj nám samotným – toto je srdce blogu *Overcoming Bias*.

Zlo sa zakráda, skryté, po neosvetlených tieňoch mysle. My sa obzeráme s jasnosťou histórie a plačeme, keď spomíname na hladomory naplánované Stalinom a Maom, ktoré zabili desiatky miliónov. Označujeme to ako zlo, pretože to bolo vykonané vedomým ľudským úmyslom spôsobiť bolesť a smrť nevinným ľudským bytostiam. Označujeme to ako zlo kvôli odporu, ktorý voči tomu cítime, keď sa obzeráme s jasnosťou histórie. Aby sa páchatelia zla vyhli tejto prirodzenej opozícii, tento odpor musí zostať skrytý. Jasnosti sa treba vyhnúť za každú cenu. Ako majú ľudia s jasným pohľadom sklon vzpierať sa zlu, ktoré vidia, tak má ľudské zlo, kdekoľvek existuje, v pláne zahmliť myslenie.

1984 to ukazuje ostro: Orwellovi najväčší zločinci sú vystrihovači a premaľovávači fotografií (založené na historickom vystrihovaní a fotografovaní v Sovietskom Zväze). A na vrchole všetkej temnoty na Ministerstve lásky, O'Brien mučí Winstona, aby pripustil, že dva plus dva sa rovná päť:[86]

> „Pamätáš sa," pokračoval, „ako si si napísal do denníka: ‚Sloboda je slobodou povedať, že dva plus dva sa rovná štyri'?"
>
> „Áno," povedal Winston.
>
> O'Brien zdvihol ľavú ruku, chrbátom k Winstonovi, so skrytým palcom a štyrmi vystretými prstami.
>
> „Koľko prstov ukazujem, Winston?"
>
> „Štyri."
>
> „A keď strana povie, že to nie sú štyri, ale päť – potom koľko?"
>
> „Štyri."
>
> Toto slovo skončilo zhíknutím od bolesti. Ručička na stupnici vyskočila na štyridsaťpäť. Po celom Winstonovom tele vyrazil pot. Vzduch sa mu dral do pľúc a potom zase von v hlbokých stonoch, ktoré ani zaťatím zubov nedokázal zastaviť. O'Brien ho sledoval, štyri prsty stále vystreté. Potiahol páku naspäť. Tentokrát bolesť ustúpila iba mierne.

Som opakovane zdesený, ako napohľad rozumní ľudia – napríklad Robinov kolega Tylen Cowen – si nemyslia, že prekonávať skreslenia je dôležité. Toto je vaša *myseľ*, o čom hovoríme. Vaša ľudská inteligencia. To, čo vás oddeľuje od opice. To, čo vybudovalo tento svet. Nemyslíte si, že je dôležité, ako myseľ funguje? Nemyslíte si, že systematické poruchy mysle sú dôležité? Myslíte si, že by Inkvizícia mučila bosorky, keby všetci boli ideálni Bayesovci?

Tyler Cowen si zrejme myslí, že prekonávanie skreslení je rovnakým skreslením ako samotné skreslenia: „Vnímam Robinov blog ako príklad skreslenia, a teda ako ukážku, že skreslenie môže byť veľmi užitočné." *Dúfam*, že je to len výsledok príliš abstraktného myslenia v snahe znieť chytro. Môže si Tyler vážne myslieť, že necitlivosť k rozsahu pri hodnote ľudského života je na rovnakej úrovni ako snaha vytvoriť plány, ktoré *naozaj* zachránia toľko životov, koľko sa len dá?

86 George Orwell, *1984* (Signet Classic, 1950).

Orwell bol nútený bojovať proti podobnému postoju – že pripustiť hocijaký rozdiel je mladistvou naivnosťou:

> Stuart Chase a ďalší už takmer tvrdia, že všetky abstraktné slová sú bezvýznamné a používajú to ako zámienku na presadzovanie istého druhu politického kvietizmu. Pokiaľ neviete, čo je to fašizmus, ako môžete bojovať proti fašizmu?

Možno prekonávanie skreslení nevyzerá dostatočne vzrušujúco, ak je podávané ako zápas s púhymi náhodnými chybami. Možno je ťažšie sa pre niečo zapáliť, ak proti nám nestojí nejaké jasné zlo. Spravme si teda absolútne jasno v tom, že kde je na tomto svete ľudské zlo, kde je krutosť a mučenie a úmyselné vraždenie, tam sú skreslenia, ktoré to zahaľujú. Kde ľudia s jasným pohľadom odporujú týmto skresleniam, zahalené zlo útočí naspäť. Pravda *má* svojich nepriateľov. Keby bol *Overcoming Bias* časopisom v bývalom Sovietskom Zväze, každý autor a diskutér na tomto blogu by bol poslaný do pracovného tábora.

V celej ľudskej histórii, každý krok dopredu bol poháňaný novou jasnosťou myslenia. Okrem niekoľkých prírodných katastrof, každá veľká pohroma bola poháňaná hlúposťou. Naším posledným nepriateľom sme my sami; toto je vojna a my sme vojaci.

* →

→ http://lesswrong.com/lw/jd/human_evil_and_muddled_thinking/

G: Proti racionalizácii

67. Vedieť o skresleniach môže ľuďom ublížiť

Jedného dňa som skúšal povedať mojej mame o probléme kalibrácie odborníkov a povedal som: „Takže keď odborník povie, že si niečím na 99 % istý, v skutočnosti sa to stáva iba v 70 % prípadov.“ Po krátkej pauze, keď som náhle uvedomil, že hovorím so svojom mamou, som rýchlo dodal: „Samozrejme, takýto skepticizmus musíš uplatňovať nestranne, aj na seba samého, nielen používať ho ako argument proti všetkému, s čím nesúhlasíš...“

A moja mama na to: „Robíš si srandu? Toto je skvelé! To budem používať vždy!“

Taberov a Lodgeov „Motivovaný skepticizmus vo vyhodnocovaní politických názorov“ popisuje potvrdenie šiestich predpovedí:[87]

1. Efekt pôvodného postoja. Osoby, ktoré majú na danú tému silný názor – aj keď sú vyzvané, aby boli objektívne – hodnotia argumenty súhlasiace s ich názorom priaznivejšie než argumenty proti.
2. Sklon nesúhlasiť. Osoby strávia viac času a kognitívnych zdrojov ohováraním argumentov proti než súhlasiacich argumentov.
3. Sklon súhlasiť. Osoby s možnosťou slobodne si vybrať svoje informačné zdroje, budú hľadať skôr súhlasiace než odporujúce zdroje.
4. **Polarizácia postoja. Vystaviť osoby napohľad vyváženej množine argumentov za a proti posilní ich pôvodný postoj.**
5. Efekt sily postoja. Osoby vyjadrujúce silnejšie postoje budú viac náchylné k horeuvedeným sklonom.
6. **Efekt sofistikovanosti. Osoby, ktoré sa vyznajú v politike, keďže majú viac zbraní na odrážanie nesúhlasiacich faktov a argumentov, budú viac náchylné k horeuvedeným sklonom.**

Ak ste na začiatku iracionálny, *viac* vedomostí vám môže *ublížiť*. Pre skutočného Bayesovca informácia nikdy nemôže mať negatívny očakávaný úžitok. Ale ľudia nie sú dokonalí bayesovskí majstri; keď si nedávame pozor, môžeme sa porezať.

Videl som ľudí vážne pošahaných vlastnou znalosťou skreslení. Mali viac zbraní na argumentovanie proti všetkému, čo sa im nepáčilo. A tento problém – príliš veľa pripravených zbraní – je jedným z hlavných dôvodov, prečo ľudia s vysokou mentálnou obratnosťou často skončia ako hlupáci, v tom zmysle hlúposti, ktorý Stanovich nazýva „dysracionalita“.

Napadajú vám ľudia, na ktorých tento popis sedí, však? Ľudia s vysokým *g*-faktorom, ktorí skončia ako *menej* efektívni, pretože sú príliš sofistikovanými diskutérmi? Myslíte si, že by ste im pomohli – urobili ich efektívnejšími racionalistami – keby ste im jednoducho povedali o zozname klasických sklonov?

Spomínam si na človeka, ktorý sa dozvedel o probléme kalibrácie / prehnanej sebadôvery. Čoskoro na to povedal: „Pozri, nemôžeš dôverovať odborníkom; ako ukazujú experimenty, veľmi často sa mýlia. Preto, keď predpovedám budúcnosť, radšej vychádzam z toho, že veci budú pokračovať tak, ako historicky...“ a pustil sa do veľmi zložitej, náchylnej k chybám, vysoko otáznej extrapolácie. Akosi, keď došlo na dôveru k jeho vlastným obľúbeným záverom, všetky tieto sklony a omyly sa zdali omnoho menej *výrazné* – napadli mu omnoho zriedkavejšie – než keď potreboval protiargument proti niekomu inému.

Povedal som mu teda o probléme so sklonom nesúhlasiť a sofistikovane argumentovať, a hľa, keď som nabudúce povedal niečo, čo sa mu nepáčilo, označil ma za sofistikovaného argumentátora. Nepokúšal sa ukázať na žiaden konkrétny sofistikovaný argument, žiadnu konkrétnu chybu – iba potriasol

87 Charles S. Taber and Milton Lodge, „Motivated Skepticism in the Evaluation of Political Beliefs,“ *American Journal of Political Science* 50, no. 3 (2006): 755–769, doi:10.1111/j.1540-5907.2006.00214.x.

hlavou a smutne si vzdychol, ako očividne používam svoju inteligenciu, aby som porazil sám seba. Získal ďalší úplne všeobecný protiargument.

Ešte aj samotný pojem „sofistikovaného argumentátora“ môže byť zhubný, ak vám príliš pohotovo prichádza na rozum, keď sa stretnete s napohľad inteligentným človekom, ktorý povie niečo, čo sa vám nepáči.

Snažím sa poučiť zo svojich chýb. Keď som mal naposledy prednášku o heuristikách a sklonoch, na začiatku som tejto všeobecný pojem ilustroval pomocou klamu konjunkcie a heuristiky reprezentatívnosti. Až potom som prešiel k sklonu súhlasiť, sklonu nesúhlasiť, sofistikovanej argumentácii, motivovanému skepticizmu, a ďalším efektom postoja. Ďalších tridsať minút som strávil *zdôrazňovaním* tejto témy, vracal som sa k nej z toľkých rôznych uhlov, ako som len vedel.

Chcel som, aby sa moje obecenstvo o túto tému začalo zaujímať. Nuž, na to by stačil jednoduchý popis klamu konjunkcie a heuristiky reprezentatívnosti. Ale predpokladajme, že sa zaujímať začali. Čo ďalej? Literatúra o sklonoch je väčšinou kognitívna psychológia pre samotnú kognitívnu psychológiu. Musel som svojmu obecenstvu dať veľké varovanie počas tejto jednej lekcie, inak by si pravdepodobne už nikdy nevypočuli.

Či v článku alebo na prednáške, odteraz sa vždy snažím nespomínať kalibráciu a prehnanú sebadôveru, dokiaľ najprv neobjasním sklon nesúhlasiť, motivovaný skepticizmus, sofistikovaných argumentátorov a dysracionalitu mentálne obratných. V prvom rade, neublížiť!

*→

—

68. Aktualizujte postupne

Politika zabíja myslenie. Debata je vojna, argumenty sú vojaci. Existuje pokušenie hľadať spôsoby, ako interpretovať všetky možné experimentálne výsledky ako potvrdzujúce vašu teóriu, ako keď opevňujete citadelu pred každým možným smerom útoku. To nemôžete. Je to matematicky nemožné. Pre každé očakávanie indície existuje rovnaké očakávanie protiindície v opačnom smere.

Ale je v poriadku, ak vaša obľúbená teória nie je *dokonale* chránená. Ak hypotéza znie, že na minci padne hlava v 95% prípadov, potom v jednom prípade z dvadsiatich uvidíte niečo, čo vyzerá ako indícia proti. To je v poriadku. To je normálne. Dokonca sa to očakáva, pokiaľ máte devätnásť potvrdzujúcich pozorovaní na každé odporujúce. Pravdepodobnostný model môže dostať jeden či dva zásahy a stále prežiť, pokiaľ tieto zásahy *nepokračujú*.

Napriek tomu sa všeobecne verí, najmä na súde verejnej mienky, že pravdivá teória nemôže mať *žiadne* zlyhania, a nepravdivá teória *žiadne* úspechy.

Nájdete ľudí, ktorí sa držia jediného kúska toho, čo považujú za indíciu, a tvrdia, že ich teória to dokáže „vysvetliť“, akoby toto bola celá podpora, akú len môže teória potrebovať. Podľa všetkého nepravdivá teória nemôže mať *žiadnu* indíciu vo svoj prospech; je nemožné, aby nepravdivá teória dokázala vysvetliť čo len jedinú udalosť. Tým pádom je jeden kus potvrdzujúcej indície všetkým, čo teória potrebuje.

Je iba o čosi menej hlúpe držať sa jediného kúska *pravdepodobnostnej* protiindície ako vyvrátenia, akoby sa proti správnej teórii nemohlo ani *trošku* argumentovať. Ale presne takto ľudia argumentovali celé veky, snažili sa poraziť všetky nepriateľské argumenty, zatiaľ čo nepriateľovi nedopriali ani jediný kúsok podpory. Ľudia chcú, aby ich debaty boli jednostranné; sú zvyknutí na svet, v ktorom proti ich obľúbeným teóriám neexistuje jediný kúsok protipodpory. Preto, pripustiť jediný kus pravdepodobnostnej protiindície by znamenalo koniec sveta.

Ja viem, že niekto v obecenstve ide povedať: „Ale v skutočnom svete *nemôžeš* pripustiť ani jediný bod, ak chceš vyhrať debatu! Ak pripustíš, že existuje nejaký protiargument, nepriateľ ti ho bude omieľať dokola a dokola – to nemôžeš nepriateľovi dovoliť! *Prehráš!* Čo by mohlo byť fyzicky desivejšie než *toto*?“

→ http://lesswrong.com/lw/he/knowing_about_biases_can_hurt_people/

No a čo. Rozumnosť nie je na vyhrávanie debát, je na rozhodovanie, ku ktorej strane sa pridať. Ak ste sa už rozhodli, za ktorú stranu idete argumentovať, rozumnú časť ste tým *ukončili*, či už dobre alebo úboho. Ale ako sa môžete, sami od seba, rozhodnúť, za ktorú stranu budete argumentovať? Ak *vybrať si nesprávnu stranu* je fyzicky desivé, hoci len trošku fyzicky desivé, mali by ste spracovať *všetky* indície.

Rozumnosť nie je chôdza, ale tanec. Pri každom kroku tohto tanca by vaša noha mala dopadnúť presne na správne miesto, nie viac doľava ani doprava. Posuňte názor vyššie s každým kúskom potvrdzujúcej indície. Posuňte názor nižšie s každým kúskom opačnej indície. Áno, *nižšie*. Aj keď je model správny, ak to nie je presný model, niekedy budete musieť prehodnotiť svoj názor smerom *nadol*.

Ak jeden či dva kúsky indície sú náhodou protipodporou vášho názoru, to je v poriadku. To sa pravdepodobnostným indíciám pre nepresné teórie občas stáva. (Keď zlyhá presná teória, vtedy *máte* problém!) Len posuňte svoj názor o čosi nadol – pravdepodobnosť, pomer šancí, alebo hoci aj neverbálnu silu dôvery vo vašej mysli. Len ho posuňte o čosi nadol a čakajte na ďalšie indície. Ak je teória pravdivá, čoskoro príde podporujúca indícia a pravdepodobnosť opäť začne šplhať nahor. Ak je teória nepravdivá, tak ju predsa naozaj nechcete.

Problém používania čierno-bieleho, binárneho, kvalitatívneho rozmýšľania je, že ľubovoľné jedno pozorovanie buď zničí teóriu alebo ju nezničí. Keď nie je dovolené jediné odporujúce pozorovanie, vytvára to kognitívnu disonanciu a treba to odargumentovať. A toto vylučuje postupný pokrok; vylučuje to správne zapojenie všetkých indícií. Pri pravdepodobnostnom uvažovaní si uvedomujeme, že správna teória bude v priemere vytvárať väčšiu váhu podpory než protipodpory. A tak si môžete *bez strachu* povedať: „Toto je mierna indícia proti, posuniem svoj názor nadol.“ Áno, *nadol*. To neničí vašu obľúbenú teóriu. To by bolo kvalitatívne uvažovanie; myslite kvantitatívne.

Pre každé očakávanie indície existuje očakávanie rovnakej protiindície v opačnom smere. V každej situácii musíte v priemere očakávať upravenie svojho názoru nadol rovnako ako musíte očakávať upravenie svojho názoru nahor. Ak si myslíte, že už viete, aká indícia príde, potom si musíte byť svojou teóriou pomerne istí – pravdepodobnosť blízka k 1 – čo vám nenecháva veľa miesta na ďalšie zvyšovanie pravdepodobnosti. A akokoľvek nepravdepodobné sa zdá, že by ste natrafili na vyvracajúcu indíciu, výsledný posun nadol musí byť dosť veľký na to, aby presne vyvážil očakávaný zisk na opačnej strane. Vážený priemer vašej očakávanej výslednej pravdepodobnosti sa musí rovnať vašej pôvodnej pravdepodobnosti.

Aké bláznivé je potom báť sa upravovania svojej pravdepodobnosti nadol, ak sa vôbec unúvate danú záležitosť vyšetrovať? V priemere musíte očakávať rovnako veľa posunu nadol ako nahor z každého jednotlivého pozorovania.

Možno sa stane, že kúsok protipodpory príde zase, a zase, a zase, zatiaľ čo nová podpora priteká pomaly. Možno zistíte, že váš názor sa posúva stále nižšie a nižšie. Až si nakoniec uvedomíte, z ktorého smeru proti vám fúka vietor indícií. V tej chvíli uvedomenia nemá zmysel zostavovať výhovorky. V tej chvíli uvedomenia ste sa svojho obľúbeného názoru *už vzdali*. Hurá! Čas na oslavu! Otvorte fľašu šampanského a pošlite pre pizzu! Držaním sa svojich pôvodných názorov sa predsa nemôžete stať silnejší.

*→

69. Jeden argument proti armáde

Včera som hovoril o spôsobe uvažovania, pri ktorom nie je dovolený jediný argument proti, s tým výsledkom, že každú nepodporujúcu indíciu treba odargumentovať. Dnes naznačujem, že keď sa ľudia stretnú s argumentom proti, zabránia zníženiu svojej istoty tak, že si *zopakujú* už známu podporu.

Predstavte si, že krajina Freedonia debatuje, či ich sused, Sylvania, môže za nedávnu hromadu meteoritov padajúcich na ich mestá. Sú niektoré indície, ktoré to naznačujú: meteority zasiahli mestá

→ http://lesswrong.com/lw/ij/update_yourself_incrementally/

blízko sylvaniánskej hranice; *pred* dopadmi bola nezvyčajná aktivita na sylvaniánskej burze cenných papierov; a sylvaniánskeho veľvyslanca Trentina niekto počul šomrať o „nebeskej pomste“.

Niekto príde a povie vám: „Nemyslím si, že Sylvania je zodpovedná za tieto padajúce meteority. Každoročne s nami obchodujú v miliardách dinárov.“ „Pozri,“ odpoviete mu, „meteority zasiahli mestá blízko Sylvanie, na ich burze cenných papierov bola podozrivá aktivita, a ich veľvyslanec potom hovoril o nebeskej pomste.“ Keďže tieto tri argumenty prevážia onen jeden, *zachováte* si názor, že za to môže Sylvania – veríte namiesto neverenia, kvalitatívne. Rovnováha indícií je zrejme proti Sylvanii.

Potom za vami príde ďalší a povie: „Nemyslím si, že Sylvania je zodpovedná za tieto padajúce meteority. Zmeniť smer padajúcich meteoritov je veľmi ťažké. Sylvania dokonca ani nemá vesmírny program.“ Odpoviete: „Ale meteority zasiahli mestá blízko Sylvanie, a ich investori to vedeli, a veľvyslanec to priamo priznal!“ Opäť, tieto tri argumenty prevážia onen jeden (prevahou troch proti jednému), takže si zachováte názor, že za to môže Sylvania.

Ba čo viac, vaše presvedčenie sa *posilnilo*. Teraz ste už v dvoch oddelených situáciách vyhodnotili rovnováhu dôkazov, a oba razy bola rovnováha naklonená proti Sylvanii s pomerom 3 ku 1.

Stretávate sa s ďalšími argumentmi prosylvanskych zradcov – opäť, a opäť, a ešte stokrát – ale zakaždým je nový argument hravo porazený 3 ku 1. A v každej situácii sa cítite viac presvedčení, že Sylvania za to naozaj môže; posúvate svoje pôvodné presvedčenie podľa pociťovanej rovnováhy indícií.

Problém je samozrejme v tom, že *opakovaním* argumentov, ktoré *ste už vedeli*, počítate danú indíciu dvojmo. To by bol vážny hriech ešte aj v prípade, že by ste dvojmo počítali *všetky* indície. (Predstavte si vedca, ktorý urobí pokus s 50 jedincami a nezíska štatisticky významné výsledky, a tak všetky údaje započíta dvakrát.)

Ale selektívne počítať dvojmo *iba niektoré indície*, to je číra fraška. Pamätám si, že ako dieťa som videl kreslený film, v ktorom záporný hrdina rozdeľoval korisť pomocou nasledujúceho algoritmu: „Jedna zlatka tebe, jedna mne. Druhá tebe, jedna-dve mne. Tretia tebe, jedna-dve-tri mne.“

Ako som v predchádzajúcej kapitole zdôraznil, aj keď je obľúbený názor *pravdivý*, racionalista môže občas potrebovať znížiť jeho pravdepodobnosť pri zapájaní *všetkých* indícií. Áno, rovnováha podpory môže stále uprednostniť váš obľúbený názor. Ale aj tak musíte posunúť pravdepodobnosť *nadol* – áno, *nadol* – z akejkoľvek hodnoty, kde bola predtým než ste počuli indíciu proti. Nepomôže *zopakovať* si podporujúce argumenty, pretože tie ste už zohľadnili.

A predsa sa mi zdá, že keď sa ľudia stretnú s *novým* protiargumentom, hľadajú ospravedlnenie, prečo neposunúť svoje presvedčenie nadol, a samozrejme nájdu podporujúce argumenty, ktoré *už poznali*. Musím si neustále dávať pozor, aby som to nerobil ja sám! Pripadá mi to rovnako prirodzené ako zablokovať úder meča pripraveným štítom.

Pri správnom druhu nesprávneho rozmýšľania, hŕstka podpory – dokonca aj jediný argument – dokáže odraziť celú armádu protirečení.

*→

70. Spodný riadok

V aukcii sú dve zapečatené krabice, krabica A a krabica B. Práve jedna z týchto krabíc obsahuje vzácny diamant. Existujú všelijaké príznaky a znamenia naznačujúce, či daná krabica obsahuje diamant; ja však o žiadnom z nich *neviem*, že by bolo celkom spoľahlivé. Napríklad na jednej krabici je modrá pečiatka, a ja viem, že krabice, ktoré obsahujú diamant, majú väčšiu šancu mať modrú pečiatku než prázdne krabice. Alebo jedna krabica má lesklý povrch a ja mám podozrenie – ale nie som si istý – že krabice s diamantom nebývajú lesklé.

Predpokladajme, že je tam chytrý rečník, ktorý drží hárok papiera a hovorí majiteľom krabice A a krabice B: „Vydražte si moje služby, a ktorý vás získa moje služby, za toho budem argumentovať,

→ http://lesswrong.com/lw/ik/one_argument_against_an_army/

že jeho krabica obsahuje diamant, takže za tú krabicu dostane vyššiu cenu.“ Takže majitelia krabíc dražia, a majiteľ krabice B ponúkne viac, čím získa služby chytrého rečníka.

Chytrý rečník si začne organizovať myšlienky. Najprv napíše: „A *preto*, krabica B obsahuje diamant!“ na spodok svojho hárku papiera. Potom na vrch papiera napíše: „Krabica B má na sebe modrú pečiatku,“ a pod to: „Krabica A je lesklá,“ a ešte: „Krabica B je ľahšia než krabica A,“ a tak ďalej, pre mnohé ďalšie príznaky a znamenia; chytrý rečník akurát ignoruje všetky znamenia, ktoré by mohli svedčiť v prospech krabice A. Nakoniec chytrý rečník príde za mnou a prečíta mi zo svojho hárku papiera: „Krabica B má na sebe modrú pečiatku, a krabica A je lesklá,“ a tak ďalej, až kým sa dostane k: „A *preto*, krabica B obsahuje diamant!“

Zvážte však: Vo chvíli, keď chytrý diskutér napísal svoj záver, vo chvíli, keď atrament vsiakol do jeho hárku papiera, previazanosť indíciami medzi týmto fyzickým atramentom a fyzickými krabicami bola pevne daná.

Možno vám pomôže predstaviť si množinu svetov – Everettove vetvy alebo Tegmarkove duplikáty – v ktorých je nejaká objektívna frekvencia, s ktorou krabica A alebo krabica B obsahuje diamant. Podobne je nejaká objektívna frekvencia v rámci podmnožiny „svety s lesklou krabicou A“, kde krabica B obsahuje diamant; a nejaká objektívna frekvencia v „svetoch s lesklou krabicou A a krabicou B s modrou pečiatkou,“ kde krabica B obsahuje diamant.

Atrament na papieri je sformovaný do čudných tvarov a kriviek, ktoré vyzerajú ako text: „A *preto*, krabica B obsahuje diamant.“ Keby ste boli gramotný anglicky hovoriaci človek, mohlo by vás to popliesť a mohli by ste si myslieť, že takto pokrútený atrament nejako *znamená*, že krabica B obsahuje diamant. Pokusné osoby, ktoré majú za úlohu povedať farbu na vytlačenom obrázku a dostanú obrázok „zelená“ (napísané červenou farbou), často povedia „zelená“ namiesto „červená“. Pomáha byť negramotný, vtedy vás nemýli tvar atramentu.

Pre nás je dôležitou stránkou veci jej previazanosť s inými vecami. Vezmime si opäť tú množinu svetov, Everettových vetiev alebo Tegmarkových duplikátov. Vo chvíli, keď všetci chytrí rečníci vo všetkých svetoch umiestnia atrament na spodok svojho papiera – predpokladajme, že je to v jednej chvíli – pevne tým určili koreláciu atramentu s krabicami. Chytrý rečník píše nezmizíkovateľným perom; atrament sa už nezmení. Krabice sa už nezmenia. V podmnožine svetov, kde atrament hovorí „A preto, krabica B obsahuje diamant,“ je už pevne dané nejaké percento svetov, kde krabica A obsahuje diamant. A to sa nezmení bez ohľadu na to, čo sa dopíše na prázdne riadky nad tým.

Takže indíciová previazanosť atramentu je pevne daná a ponechávam na vás, aby ste sa rozhodli, aká môže byť. Možno majitelia krabíc, ktorí veria, že sa v ich prospech dá zostaviť lepší argument, sú ochotnejší prenajať si reklamu; možno majitelia krabíc, ktorí sa boja svojej nedostatočnosti, dražia vyššie. Ak majitelia krabíc sami nerozumejú príznakom a znameniam, potom bude atrament naprosto nepreviazaný s obsahmi krabíc, hoci vám môže povedať niečo o majiteľových financiách a dražiacich zvykoch.

Teraz si predstavte, že iný človek je úprimne zvedavý, a *najprv* si na papier napíše všetky odlišné príznaky *oboch* krabíc, potom aplikuje svoje poznanie a zákony pravdepodobnosti, a napíše naspodok: „*Preto* odhadujem s pravdepodobnosťou 85 %, že krabica B obsahuje diamant.“ Čoho indíciou je tento nápis? Keď skúmam reťaz príčiny a následku vedúcu k tomuto fyzickému atramentu na fyzickom papieri, zisťujem, že reťaz kauzality sa vinie cez všetky príznaky a znamenia krabíc, a závisí od týchto príznakov; vo svetoch s odlišnými príznakmi je naspodku papiera napísaná iná pravdepodobnosť.

Preto je písmo zvedavého výskumníka previazané s príznakmi a znameniami a s obsahom krabíc, zatiaľ čo písmo chytrého rečníka je indíciou iba toho, ktorý majiteľ mu v dražbe ponúkol viac. Toto je veľký rozdiel vo význame atramentu, hoci ten, kto hlúpo nahlas prečíta tvar atramentu, si môže myslieť, že napísané slová znejú podobne.

Vaša efektivita ako racionalistu je daný tým, ktorý algoritmus naozaj napíše spodný riadok vašich myšlienok. Ak vaše auto vydáva pri brzdení kovové škrípajúce zvuky, ale vy nie ste ochotný čeliť finančným nákladom na opravu brzd, môžete sa rozhodnúť hľadať dôvody, prečo vaše auto netreba opraviť. Ale skutočné percento vás, ktoré prežije v Everettových vetvách alebo Tegmarkových svetoch –

ktoré použijeme na opis vaše efektivity ako racionalistu – je dané algoritmom, ktorý rozhodne, pre *ktorý* záver budete hľadať argumenty. V tomto prípade je skutočný algoritmus: „Nikdy neopravuj nič drahé.“ Ak je to dobrý algoritmus, fajn; ak je to zlý algoritmus, smola. Argumenty, ktoré dopíšete potom, nad spodný riadok, nezmenia nič žiadnym smerom.

Toto je myslené ako varovanie proti vášmu vlastnému uvažovaniu, nie ako Úplne Všeobecný Protiargument proti záverom, ktoré sa vám nepáčia. Pretože povedať „Môj oponent je chytrý rečník“ je chytrý argument, ak platíte sami sebe, aby ste si zachovali všetky názory, ktoré ste mali na začiatku. Najchytrejší rečník na svete môže poukázať na to, že slnko svieti, a predsa bude naďalej deň.

*→

71. Filtrovaná indícia

Včera som rozoberal dilemu chytrého rečníka, najatého, aby vám predal krabicu, ktorá možno obsahuje diamant a možno nie. Chytrý rečník vám zdôrazní, že na krabici je modrá pečiatka, a že je známy fakt, že krabice obsahujúce diamant majú väčšiu šancu mať modrú pečiatku než prázdne krabice. Čo sa v tomto bode deje z bayesovského pohľadu? Musíte bezmocne aktualizovať svoje pravdepodobnosti tak, ako si to chytrý rečník želá?

Ak sa môžete pozrieť na danú krabicu osobne, môžete si všetky príznaky zrátať sami. Čo ak sa nemôžete pozrieť? Čo ak jediná indícia, ktorú máte, je slovo chytrého rečníka, ktorý je zákonom obmedzený na hovorenie iba pravdivých výrokov, ale nehovorí vám všetko, čo vie? Každý výrok, ktorý povie, je platná indícia – ako by ste mohli *neaktualizovať* svoje pravdepodobnosti? Prestalo azda byť pravdou, že v takom a takom percente Everettových vetiev alebo Tegmarkových duplikátov, v ktorých krabica B má modrú pečiatku, krabica B obsahuje diamant? Podľa Jaynesa, Bayesovec musí vždy zohľadniť všetky známe indície, pod hrozbou paradoxu. Lenže potom vás chytrý rečník dokáže presvedčiť o hocičom, čo bude chcieť, pokiaľ je dostatočná paleta príznakov, o ktorých vás môže selektívne informovať. To neznie správne.

Vezmite si jednoduchý prípad, nevyváženú mincu, na ktorej buď padá 2/3 hlava a 1/3 znak, alebo 1/3 hlava a 2/3 znak; obe tieto možnosti sú a priori rovnako pravdepodobné. Každé pozorované H je 1 bit indície pre H-nevyváženú mincu; každé pozorované Z je jeden bit indície pre Z-nevyváženú mincu. Hodím mincou desaťkrát a potom vám poviem: „Pri 4. hode, 6. hode, a 9. hode padla hlava.“ Aká je vaša výsledná pravdepodobnosť, že minca je H-nevyvážená.

A odpoveď je, že to môže byť takmer čokoľvek, podľa toho, aká reťaz príčin a následkov leží za mojím vyslovením týchto slov – mojím výberom, ktoré hody vám oznámiť.

- Možno sa riadim algoritmom, že poviem výsledok 4., 6. a 9. hodu bez ohľadu na výsledky, ktoré padli v týchto alebo ostatných hodoch. Ak viete, že som použil tento algoritmus, výsledné šance sú 8:1 v prospech H-nevyváženej mince.
- Možno oznamujem všetky hody, na ktorých padla hlava, a nič iné. V tom prípade viete, že vo všetkých zvyšných 7 hodoch padol znak, a výsledné šance sú 1:16 proti H-nevyváženej minci.
- Možno som sa dopredu rozhodol, že vám poviem výsledok 4., 6. a 9. hodu iba vtedy, ak pravdepodobnosť, že minca je H-nevyvážená prekročí 98 %. A tak ďalej.

Alebo si vezmite problém Montyho Halla:

> V hernej show dostanete na výber tri dvere vedúce do troch miestností. Viete, že v jednej miestnosti je 100 000 dolárov, a zvyšné dve sú prázdne. Moderátor vás požiada, aby ste si vybrali dvere, a vy si vyberiete dvere číslo 1. Potom moderátor otvorí dvere číslo 2 a ukáže prázdnu izbu. Chcete zameniť svoje rozhodnutie za číslo 3 alebo zostávate s dverami číslo 1?

→ http://www.lesswrong.com/lw/js/the_bottom_line/

Odpoveď závisí od moderátorovho algoritmu. Ak moderátor vždy otvorí niektoré dvere a vždy si vyberie dvere vedúce do prázdnej miestnosti, potom by ste mali zameniť rozhodnutie za dvere číslo 3. Ak moderátor vždy otvára dvere číslo 2 bez ohľadu na to, čo je za nimi, dvere 1 aj 3 majú obe pravdepodobnosť 50 %, že obsahujú peniaze. Ak moderátor otvára nejaké dvere iba v tých prípadoch, keď ste si na prvýkrát vybrali dvere s peniazmi, potom by ste mali jednoznačne zostať s číslom 1.

Nemali by ste zohľadniť iba to, že dvere číslo 2 sú prázdne, ale tento fakt plus fakt, že sa moderátor *rozhodol* otvoriť dvere číslo 2. Mnohí ľudia sú zo štandardného problému Montyho Halla zmätení, pretože aktualizujú len o informáciu, že číslo 2 je prázdne, a v tom prípade majú čísla 1 a 3 rovnakú pravdepodobnosť obsahovať peniaze. To je dôvod, prečo majú Bayesovci prikázané zohľadňovať všetky svoje vedomosti, pod hrozbou paradoxu.

Keď niekto povie: „Pri 4. hode mincou padla hlava," nezohľadňujeme iba to, že pri 4. hode mincou padla hlava – neberieme si podmnožinu všetkých možných svetov, kde v 4. hode padla hlava – ale zohľadňujeme podmnožinu všetkých možných svetov, kde rečník riadiaci sa nejakým konkrétnym algoritmom *povedal*: „Pri 4. hode mincou padla hlava." Vyslovená veta nie je sama osebe faktom; nenechajte sa zviesť z cesty púhym významom slov.

Väčšina súdnych procesov funguje podľa teórie, že každý prípad má práve dve odporujúce si strany, a že je jednoduchšie nájsť dvoch zaujatých ľudí než jedného nezaujatého. Či už žalobca alebo obhajca, *niekto* z nich má motív predložiť každý kúsok indície, takže súd uvidí všetky indície; to je teória. Ak sú v dileme s dvoma krabicami dvaja chytrí rečníci, nie je to až také dobré ako jeden zvedavý výskumník, ale je to skoro také dobré. Ale to platí, keď sú dve krabice. Skutočnosť často obsahuje mnohostranné problémy, a hlboké problémy, a nesamozrejmé odpovede, ktoré sa nenájdu ľahko tým, že Modrí a Zelení kričia jeden na druhého.

Dajte si pozor, aby ste nezneužívali pojem filtrovania indície ako Úplne Všeobecný Protiargument na vylúčenie všetkých indícií, ktoré sa vám nepáčia: „Tento argument bol filtrovaný, preto ho môžem ignorovať." Ak vás podráždil argument opačnej strany, potom ste už zoznámení s prípadom, a dosť zaujatí na to, aby ste si vybrali jednu stranu. Pravdepodobne už poznáte najsilnejšie argumenty vašej strany. Nemáte žiaden dôvod dedukovať na základe argumentu proti, že existujú nové príznaky a znamenia vo váš prospech, ktoré ste ešte nevideli. Zostávajú vám teda samotné nepohodlné fakty; modrá pečiatka na krabici B je stále indíciou.

Ale ak počujete daný argument po prvýkrát, a ak počujete iba jednu stranu argumentu, potom by ste si veru mali dať pozor! V istom zmysle, nikto nemôže *naozaj* veriť teórii prirodzeného výberu, dokiaľ nepočúval kreacionistov aspoň päť minút; až *potom* vie, že je to nabetón.

*→

72. Racionalizovanie

V článku „Spodný riadok" som predložil dilemu dvoch krabíc, z ktorých iba jedna obsahuje diamant, s rôznymi príznakmi a znameniami ako indíciami. Definoval som rozdiel medzi zvedavým výskumníkom a chytrým rečníkom. Zvedavý výskumník si zapíše všetky príznaky a znamenia, spracuje ich, a nakoniec napíše: „*Preto* odhadujem, že s pravdepodobnosťou 85% krabica B obsahuje diamant." Chytrý rečník pracuje pre toho, kto si ho vydražil, a začne tým, že napíše: „*Preto*, krabica B obsahuje diamant," a potom vyberá priaznivé príznaky a znamenia do riadkov nad tým.

Ten prvý postup je rozumnosť. Ten druhý postup je všeobecne známy ako „racionalizovanie".

„Racionalizovanie." Aký čudný názov. Podľa mňa je to *nesprávne slovo*. Nemôžete „racionalizovať" niečo, čo nie je už racionálne. To je ako keby sme „klamanie" nazývali „pravdizovanie".

Na čisto výpočtovej úrovni je obrovský rozdiel medzi:

→ http://www.lesswrong.com/lw/jt/what_evidence_filtered_evidence/

1. Začatím od indícií, a potom spracovaním tokov pravdepodobnosti, s cieľom vypísať pravdepodobný záver. (Zapísať si všetky príznaky a znamenia, a potom sa nechať unášať k pravdepodobnosti na spodnom riadku, ktorá závisí od týchto príznakov a znamení.)
2. Začatím od záveru, a potom spracovaním tokov pravdepodobnosti s cieľom vypísať indície, ktoré napohľad podporujú tento záver. (Zapísať si spodný riadok, a potom plávať proti prúdu a vyberať príznaky a znamenia na prezentáciu v horných riadkoch.)

Ktorý blázon vymyslel také mätúco podobné slová „racionalita“ a „racionalizovanie“, aby opísal takéto mimoriadne odlišné myšlienkové procesy? Dal by som prednosť pojmom, ktoré by urobili tento algoritmický rozdiel zrejmým, povedzme „rozumnosť“ verzus „obrovská vsávajúca kognitívna čierna diera“.

Nie každá zmena je zlepšenie, ale každé zlepšenie je nevyhnutne zmena. Nemôžete pre pevne daný výrok získať viac pravdy tým, že preň budete argumentovať; môžete o ňom presvedčiť viac ľudí, ale nemôžete ho urobiť *pravdivejším*. Aby sme zlepšili svoje názory, nevyhnutne musíme meniť svoje názory. Rozumnosť je operácia, ktorú používame na získanie väčšej pravdivostnej hodnoty pre naše názory tým, že ich meníme. Racionalizovanie funguje, aby upevnilo názory; lepšie by ho bolo nazvať „anti-rozumnosť“, aj kvôli praktickým výsledkom, aj kvôli otočenému algoritmu.

„Rozumnosť“ je plávanie *po prúde*, ktoré zbiera indície, zvažuje ich, a vypisuje záver. Zvedavý výskumník používa algoritmus plávania po prúde: *najprv* zbiera indície, píše si zoznam všetkých viditeľných príznakov a znamení, ktoré potom *po prúde* spracuje, aby získal predtým neznámu pravdepodobnosť, že krabica obsahuje diamant. Po celý ten čas proces rozumnosti šiel dopredu, zvedavý výskumník ešte nepoznal svoj cieľ, a práve preto bol *zvedavý*. Pri Bayesovej Ceste sa pôvodná pravdepodobnosť rovná očakávanej výslednej pravdepodobnosti. Ak poznáte svoj cieľ, už ste tam.

„Racionalizácia“ je plávanie *proti prúdu*, od záveru k vyberaným indíciám. Najprv si napíšete spodný riadok, ktorá je známy a pevne daný; cieľom vášho spracovávania je zistiť, ktoré argumenty máte napísať na predchádzajúce riadky. Toto, a nie záver, je neznámou premennou daného procesu.

Obávam sa, že Tradičná Rozumnosť nerobí svojich používateľov primerane citlivými na rozdiely medzi plávaním po prúde a proti prúdu. V Tradičnej Rozumnosti nie je nič zlé na tom, keď si vedec nájde svoju obľúbenú hypotézu a potom začne hľadať pokus, ktorý ju dokáže. Tradičný Racionalista by sa na toto pozrel súhlasne a povedal: „Pýcha je motorom, ktorý poháňa Vedu vpred.“ Áno, *je* to motor, ktorý poháňa Vedu vpred. Ľahšie nájdete žalobcu a obhajcu zaujatých opačnými smermi, než jedného nezaujatého človeka.

Ale že niečo robí každý, to ešte danú vec nerobí správnou. Bolo by ešte lepšie, keby ten vedec, keď nájde svoju obľúbenú hypotézu, ju začal *testovať* z čírej *zvedavosti* – vytvárať pokusy, ktoré budú posúvať jeho vlastné názory neznámym smerom.

Ak naozaj neviete, kam idete, pravdepodobne ste celkom zvedaví, kam to bude. Zvedavosť je prvou cnosťou, bez ktorej vaše otázky budú bezcieľne a vaše schopnosti nesústredené.

Pociťujte prúd Sily, a dajte si pozor, aby neprúdila naopak.

*→

73. Rozumný argument

Vaše povolanie je menežer kampaní a práve vás najal Mortimer Q. Snodgrass, kandidát Zelených na starostu Hadleyburgu. Ako menežer kampane si čítate blog o rozumnosti a po rozume vám behá najmä jedna myšlienka: „Ako môžem zostaviť bezchybne rozumný argument, že Mortimer Q. Snodgrass je najlepší kandidát na starostu Hadleyburgu?“

Prepáčte. To sa nedá.

→ http://www.lesswrong.com/lw/ju/rationalization/

„Čože?“ skríknete. „Ale čo ak použijem iba platnú podporu na konštrukciu svojej štruktúry rozmýšľania? Čo ak každý citovaný fakt bude pravdivý podľa môjho najlepšieho vedomia, a bude to relevantná indícia podľa Bayesovho pravidla?“

Prepáčte. Ani tak sa to nedá. Porazili ste sám seba v tom okamihu, keď ste si dopredu určili záver vášho argumentu.

Tento rok *Hadleyburský Hlásnik* poslal dotazník so 16 položkami všetkým kandidátom na starostu, s otázkami ako: „Dokážete namaľovať všetky farby vetra?“ a „Nadychujete sa?“ Žiaľ, budovu *Hlásnika* zničil meteorit pred uverejnením výsledkov. Čo je škoda, pretože váš kandidát, Mortimer Q. Snodgrass, vychádza dobre v porovnaní so svojimi súpermi v 15 zo 16 otázok. Jediná problematická časť bola otázka číslo 11: „Ste teraz, alebo ste niekedy v minulosti boli, superzloduch?“

Ste teda v pokušení uverejniť tento dotazník ako súčasť svojej kampaňovej literatúry... samozrejme s vynechaním otázky číslo 11.

Čím ste prekročili hranicu medzi *rozumnosťou* a *racionalizáciou*. Už nie je možné, aby sa voliči rozhodovali iba na základe faktov; musia sa rozhodovať aj na základe doplňujúceho faktu ich prezentácie, a odvodzovať existenciu skrytých indícií.

Tú hranicu ste vlastne prekročili v bode, keď ste zvažovali, či tento dotazník je pre vášho kandidáta priaznivý alebo nepriaznivý pred tým, ako ste sa rozhodli, či ho uverejníte. „Čože!“ vykríknete. „Kampaň by mala uverejňovať fakty nepriaznivé pre svojho kandidáta?“ Ale postavte sa do role voliča, ktorý si stále skúša vybrať kandidáta – prečo by ste cenzurovali užitočnú informáciu? Keby ste boli úprimne zvedaví, nerobili by ste to. Keby ste plávali *po prúde* od indícií k neznámej voľbe kandidáta, namiesto plávania *proti prúdu* od vopred určeného kandidáta smerom k výberu argumentov.

„Logický“ argument je taký, ktorý vyplýva zo svojich predpokladov. Napríklad nasledujúci argument je *nelogický*:

- Všetky obdĺžniky majú štyri strany.
- Všetky štvorce majú štyri strany.
- *Preto* všetky štvorce sú obdĺžniky.

Tento sylogizmus nezachráni pred nelogickosťou pravdivosť jeho predpokladov, dokonca ani pravdivosť jeho záveru. Oplatí sa odlišovať logické úsudky od nelogických, a odmietať ich ospravedlňovanie dokonca aj v prípade, že je záver zhodou okolností pravdivý. Po prvé, toto rozlišovanie môže ovplyvniť, ako budeme prehodnocovať svoje názory vo svetle budúcich indícií. Po druhé, lajdáctvo je návykové.

Hlavne ide o to, že tento sylogizmus nepredkladá skutočné vysvetlenie. Možno všetky štvorce sú obdĺžniky, ale ak áno, nie je to *preto*, že jedny aj druhé majú štyri strany. Mohli by sme to označiť ako pokrytecký sylogizmus – taký, v ktorom je nesúlad medzi jeho uvedenými dôvodmi a skutočnými dôvodmi.

Ak naozaj chcete predložiť čestný, rozumný argument *za vášho kandidáta* v politickej kampani, existuje iba jeden spôsob, ako to urobiť:

- *Skôr než vás niekto najme*, zhromaždite si všetky dostupné indície o rôznych kandidátoch.
- Urobte si zoznam bodov, ktoré použijete pri rozhodovaní, ktorý kandidát vyzerá najlepšie.
- Spracujte tento zoznam.
- Choďte za vyhrávajúcim kandidátom.
- Ponúknite mu, že sa stanete menežerom jeho kampane.
- Keď sa vás opýtajú na kampaňovú literatúru, vytlačte svoj zoznam.

Iba takto dokážete ponúknuť *rozumnú* reťaz argumentov, ktorej spodný riadok bol napísaný smerom *po prúde* na základe riadkov napísaných nad ním. To, čo *naozaj* rozhoduje o vašom spodnom riadku, je jediná vec, ktorú môžete *poctivo* napísať do riadkov nad ním.

*→

→ http://lesswrong.com/lw/jw/a_rational_argument/

74. Vyhýbanie sa naozaj slabým miestam vašich názorov

Pred pár rokmi zomrela moja prababička, vyše deväťdesiatročná, po dlhom, pomalom a krutom rozpade. Nikdy som ju nepoznal ako osobu, ale v mojom vzdialenom detstve varievala pre svoju rodinu; pamätám si jej fašírované ryby, a jej tvár, a že bola ku mne milá. Na jej pohrebe rečnil môj prastrýc, ktorý sa o ňu staral počas jej posledných rokoch: povedal, prehĺtajúc slzy, že Boh si jeho matku vzal naspäť po kúskoch: jej pamäť, jej reč, a nakoniec jej úsmev; a že keď si Boh nakoniec vzal jej úsmev, vedel, že už nebude dlho trvať kým zomrie, pretože to znamenalo, že už je takmer celá preč.

Počul som to a bol som v rozpakoch, pretože bola nemysliteľne krutá vec urobiť toto *komukoľvek*, a preto som nečakal, že to môj prastrýc pripíše Bohu. Zvyčajne, žid akosi skrátka-nemyslí-na logické dôsledky toho, že Boh dopustil tragédiu. Podľa židovskej teológie, Boh nepretržite udržiava celý vesmír a vyberá každú jednu udalosť v ňom; ale bežne je odvodzovanie logických dôsledkov z tohto presvedčenia vyhradené na šťastné udalosti. Keď poviete „to urobil Boh!“ iba keď ste požehnaní bábätkom, ale skrátka-nemyslíte-na „to urobil Boh!“ pri potratoch a úmrtiach v kolíske, dokážete si vybudovať dosť pokrivený obraz o láskavej povahe vášho Boha.

Preto som bol prekvapený, keď som počul môjho prastrýca prisudzovať pomalý rozpad svojej matky vedomému, strategicky naplánovanému Božiemu činu. Porušovalo to pravidlá náboženského sebaklamu, ako som ich ja chápal.

Keby som si bol všimol svoj vlastný zmätok, mohol som urobiť úspešnú prekvapujúcu predpoveď. Onedlho na to sa môj prastrýc vzdal židovského náboženstva. (Ako jediný člen mojej širokej rodiny okrem mňa, pokiaľ viem.)

Moderný ortodoxný judaizmus sa nepodobá na žiadne iné náboženstvo, o ktorom som počul, a neviem ako ho opísať niekomu, kto nebol nútený študovať Mišnu a Gemaru. Je v ňom tradícia pochybovania, ale ten *druh* pochybovania... Nebolo by vôbec prekvapujúce počuť rabína pri týždennej kázni, ktorý by poukázal na konflikt medzi siedmimi dňami stvorenia a 13,7 miliardami rokov od Veľkého Tresku – pretože by si myslel, že pre to našiel naozaj chytré vysvetlenie, zahŕňajúce tri ďalšie biblické odkazy, Midraš, a napoly pochopený článok zo *Scientific American*. V ortodoxnom judaizme máte povolené všimnúť si nezrovnalosti a protirečenia, ale iba za účelom ich odkecania, a kto príde s najkomplikovanejším vysvetlením, dostane odmenu.

Je tam tradícia skúmania. Ale útočíte na ciele iba so zámerom obrániť ich. Útočíte iba na tie ciele, o ktorých viete, že ich obránite.

V modernom ortodoxnom judaizme som nepočul veľké zdôrazňovanie cnosti slepej viery. Máte dovolené pochybovať. Akurát nemáte dovolené *úspešne* pochybovať.

Predpokladám, že veľká väčšina vzdelaných ortodoxných židov v niektorom bode svojho života pochybovala o svojej viere. Ale to pochybovanie bolo pravdepodobne niečo ako toto: „Podľa skeptikov, Tóra hovorí, že vesmír bol stvorený za sedem dní, čo nie je vedecky presné. Ale boli by pôvodní príslušníci Izraela, zhromaždení pri hore Sinaj, schopní pochopiť vedeckú pravdu, aj keby im bola predložená? Mali vôbec slovo ‚miliarda‘? Je jednoduchšie vidieť sedemdňový príbeh ako metaforu – najprv Boh stvoril svetlo, čo predstavuje Veľký Tresk...“

Je toto ten najslabší bod, v ktorom má človek napadnúť svoj vlastný judaizmus? Čítajte v Tóre o kúsok ďalej a nájdete, ako Boh zabíja deti, prvorodených synov v Egypte, aby presvedčil nevoleného faraóna, aby prepustil otrokov, ktorý logicky mohli byť z krajiny odteleportovaní. Ortodoxný žid túto pasáž celkom iste pozná, pretože by si mal raz za rok prečítať celú Tóru v synagóge, a s touto udalosťou sa spája veľký sviatok. Meno „Pascha“ [Veľká Noc; po anglicky Passover] pochádza z toho, že Boh *obišiel* [po anglicky: pass over] židovské domácnosti, keď zabíjal každého prvorodeného syna v Egypte.

Moderní ortodoxní židia sú, prevažne, milí a civilizovaní ľudia; omnoho civilizovanejší než tých niekoľko redaktorov Starého Zákona. Ešte aj starí rabíni boli civilizovanejší. V Sederi je rituál, kde vezmete desať kvapiek vína zo svojho pohára, jednu kvapku za každú z desiatich pohrôm, aby ste zdôraznili utrpenie Egypťanov. (Samozrejme, že by ste mali súcitiť s utrpením Egypťanov, ale nie *natoľko* súcitiť, aby by ste vstali a povedali: „Toto nie je správne! Je *zlé* urobiť niečo také!“) Ukazuje to

zaujímavý kontrast – rabíni boli podstatne láskavejší než redaktori Starého Zákona, keďže chápali krutosť pohrôm. Lenže Veda bola za oných dní slabšia, takže rabíni mohli uvažovať nad menej príjemnými stránkami Písma bez obáv, že by to celkom zlomilo ich vieru.

Keď sa ani *neopýtate*, či táto udalosť nevrhá na Boha zlé svetlo, nie ste nútení rýchlo vyhŕknuť: „Cesty Božie sú tajomné!“ alebo „Nie sme dosť múdri na to, aby sme spochybňovali Božie rozhodnutia!“ alebo „Zabíjať bábätká je okej, keď to robí Boh!“ Na túto časť otázky skrátka-nemyslíte.

Dôvod, prečo vzdelaní ľudia zostanú veriacimi, je podľa mňa to, že keď pochybujú, podvedome sú veľmi opatrní, aby napádali svoje vlastné presvedčenie iba na jeho najsilnejších miestach – na miestach, o ktorých vedia, že ich obránia. Navyše, na miestach, kde precvičenie si štandardnej obrany pôsobí posilňujúco.

Pravdepodobne pôsobí naozaj dobre precvičiť si, napríklad, svoju predpísanú obranu: „Nehovorí azda Veda, že vesmír sú atómy, ktoré sa iba nezmyselne od seba odrážajú?“, pretože potvrdzujú zmysel vesmíru a ako pochádza od Boha, atď. Omnoho príjemnejšia téma na rozmýšľanie, než negramotná egyptská matka bedákajúca nad kolískou svojho zamordovaného syna. Každý, kto *spontánne* pomyslí na to druhé, zatiaľ čo pochybuje o svojej viere v judaizmus, ten o nej *naozaj* pochybuje, a pravdepodobne už dlho nezostane židom.

Mojou pointou teraz nie je mlátiť do ortodoxného judaizmu. Som si istý, že na zabíjanie prvorodených existuje taká či onaká odpoveď, pravdepodobne celý tucet. Mojou pointou je, že keď ide o spontánne sebaspochybňovanie, človek má sklon spontánne zaútočiť na svoje silné stránky, kde si môže precvičiť upokojujúcu odpoveď, než spontánne zaútočiť na tie najslabšie, najzraniteľnejšie body. Podobne, človek má sklon zastaviť sa pri prvej odpovedi a upokojiť sa, namiesto ďalšieho kritizovania odpovede. Lepší nadpis než „Vyhýbanie sa naozaj slabým miestam vašich názorov“ by bol „Spontánne nemyslenie na najboľavejšie slabé miesta vašich názorov“.

Viac než čokoľvek iné, vplyv náboženstva udržiavajú ľudia, ktorí skrátka-nemyslia-na skutočné slabé miesta svojho náboženstva. Nemyslím si, že je to vecou tréningu, ale vecou inštinktu. Ľudia nemyslia na skutočné slabé miesta svojich vlastných názorov z rovnakého dôvodu, ako sa nedotýkajú do červena rozpálenej pece; pretože to *bolí*.

Aby ste to robili lepšie: Keď pochybujete o niektorom zo svojich najobľúbenejších názorov, zavrite oči, vyprázdnite si myseľ, zatnite zuby, a úmyselne myslite na to, čo najviac bolí. Neprecvičujte si štandardné námietky, ktorých štandardné obrany vám prinášajú dobrý pocit. Opýtajte sa sami seba, čo by *múdri* ľudia, ktorí nesúhlasia, povedali na vašu prvú odpoveď, aj na vašu druhú odpoveď. Kedykoľvek sa pristihnete, že cúvate pred námietkou, na ktorú ste letmo pomysleli, vytiahnite ju do popredia svojej mysle. Udrite sami seba na solar plexus. Pichnite si nôž do srdca a pokývajte ním, aby ste rozšírili dieru. Zoči-voči bolesti si precvičujte iba toto:

> Čo je pravda, to už je pravda.
>
> Priznať si to, to nerobí o nič horším.
>
> Nebyť voči tomu otvorený nespôsobí, že to odíde.
>
> A pretože je to pravda, je to to, s čím treba interagovať.
>
> Čokoľvek nepravdivé sa nedá prežívať.
>
> Ľudia dokážu zvládnuť to, čo je pravda,
>
> pretože to už zvládajú.
>
> --Eugene Gendlin[88]

*→

88 Eugene T. Gendlin, *Focusing* (Bantam Books, 1982).

→ http://lesswrong.com/lw/jy/avoiding_your_beliefs_real_weak_points/

75. Motivované zastavenie a motivované pokračovanie

Aj keď nesúhlasím s niektorými názormi skupiny Rýchlo a šetrne – myslím si, že sa snažia premeniť na limonádu *príliš* veľa citrónov – zároveň sa mi zdá, že majú sklon vyvinúť si *psychologicky najrealistickejšie* modely zo všetkých škôl rozhodovania. Väčšina pokusov predkladá pokusným osobám možnosti, pokusná osoba si vyberie možnosť, a to je výsledok pokusu. Zástancovia šetrnosti si uvedomujú, že v skutočnom živote si musíte *vytvárať* vlastné možnosti, a študovali, ako pokusné osoby robia *toto*.

Podobne mnohé pokusy predkladajú indície na striebornom podnose, ale v skutočnom živote musíte indície zbierať, čo môže byť drahé, a v niektorom bude sa musíte rozhodnúť, že máte dosť indícií, aby ste sa zastavili a vyberali. Keď si idete kúpiť dom, nemáte na výber presne 10 domov a nikto vás nevedie na prehliadku všetkých z nich skôr než sa môžete rozhodnúť. Pozriete sa na jeden dom, na druhý, porovnáte ich navzájom; prispôsobíte svoje očakávania – prehodnotíte nakoľko naozaj chcete byť blízko svojho pracoviska a koľko ste naozaj ochotní zaplatiť; rozhodnete sa, ktorý dom si pozriete potom, a v niektorom bode sa rozhodnete, že už ste videli domov dosť a vyberiete si.

Gilovichovo rozlišovanie medzi *motivovaným skepticizmom* a *motivovanou dôverčivosťou* zdôrazňuje, ako závery, ktorým daná osoba nechce veriť, posudzuje podľa prísnejšieho štandardu ako závery, ktorým chce veriť. Motivovaný skeptik sa pýta, či ho indície *nútia* prijať daný záver; motivovaný dôverčivec sa pýta, či mu indície *dovoľujú* prijať daný záver.

Tvrdím, že analogické skreslenie v psychologicky realistickom hľadaní je *motivované zastavenie* a *motivované pokračovanie*: keď máme *skrytý* motív vybrať si momentálne „najlepšiu“ možnosť, máme skrytý motív zastaviť sa a vybrať si, a odmietnuť zvažovanie prípadných ďalších možností. Ak máme skrytý motív odmietnuť momentálne najlepšiu možnosť, máme skrytý motív počkať s rozhodnutím až do ďalších indícií, aby sme vytvorili viac možností – aby sme našli niečo iné, hocičo, čo by sme mohli urobiť *namiesto* daného záveru.

Veľkým historickým škandálom v štatistike bol R. A. Fisher, významný zakladateľ odvetvia, ktorý trval na tom, že sa nepodarilo dokázať *kauzálne* spojenie medzi fajčením a rakovinou pľúc. „Korelácia nie je kauzalita,“ vypovedal pred kongresom. Možno fajčiari majú gén, ktorý zároveň spôsobuje, že fajčia, aj že dostanú rakovinu pľúc.

Alebo možno fakt, že Fisher bol zamestnaný ako konzultant pre tabakové firmy, mu dal skrytý motív rozhodnúť sa, že doteraz zhromaždené indície sú nedostatočné na vyvodenie záveru, a že je lepšie hľadať ďalej. Fisher bol sám fajčiar a zomrel na rakovinu hrubého čreva v roku 1962.

(Poznámka ad hominem: Fisher bol frekventista. Bayesovci[→] sú rozumnejší pri usudzovaní o pravdepodobnej kauzalite.)

Tak ako mnohé iné formy motivovaného skepticizmu, motivované pokračovanie sa môže maskovať za cnostnú rozumnosť. Kto už môže argumentovať proti zhromažďovaniu ďalších indícií?[→] Ja. Indície sú často drahé, a ešte horšie, pomalé, a určite nie je nič cnostné na odmietaní spracovať tie indície, ktoré už máte. Vždy si to môžete neskôr rozmyslieť. (Zdanlivý rozpor sa rieši takto: Stráviť *jednu hodinu* diskutovaním problému, keď máte myseľ starostlivo zbavenú riešení, je niečo iné ako čakať na ďalšiu štúdiu za 20 miliónov dolárov.)

Motivované zastavenie sa vyskytuje všade, kde sa bojíme tretej možnosti, a kde máme argument, ktorého zrejmý protiargument by sme radšej nevideli, a na ďalších miestach. Vyskytuje sa, keď hľadáte plán činnosti, ktorý vám dá dobrý pocit zo samotnej činnosti[→] a tak by ste radšej neriešili, ako dobre váš plán *naozaj* funguje, zo strachu, že zničíte hrejivú žiaru morálneho uspokojenia, za ktorú ste zaplatili kopec peňazí. Vyskytuje sa, keď sa vaše názory a očakávania rozsynchronizujú, takže máte dôvod báť sa zhromažďovania nových indícií.

→ http://bayes.cs.ucla.edu/BOOK-2K/

→ http://www.overcomingbias.com/2007/01/conspicuous_con.html

→ http://lesswrong.com/lw/kb/cant_say_no_spending/

Ponaučenie je, že samotné rozhodnutie ukončiť proces hľadania (dočasne alebo natrvalo) je rovnako ako samotný proces hľadania predmetom skreslení a skrytých motívov. Mali by ste mať podozrenie na motivované zastavenie, keď ukončíte hľadanie po tom, čo nájdete pohodlný záver, a predsa je veľa rýchlych a lacných indícií, ktoré ste ešte nezhromaždili - webové stránky, ktoré ste mohli navštíviť, protiargumenty na protiargumenty, ktoré ste mohli zvážiť, alebo ste nezavreli oči na päť minút podľa hodín, aby ste skúsili vymyslieť lepšiu možnosť. Mali by ste mať podozrenie na motivované pokračovanie, keď nejaká indícia ukazuje smerom, ktorý sa vám nepáči, ale rozhodnete sa, že potrebujete viac indícií – *drahých* indícií, o ktorých viete, že ich tak rýchlo nezhromaždíte, v porovnaní s tým, čo si nájdete na Googli za 30 minút – skôr než urobíte niečo nepohodlné.

*→

76. Falošné zdôvodnenie

Mnohí kresťania, ktorí prestali naozaj veriť, teraz tvrdia, že si vážia Bibliu ako zdroj etických rád. Štandardnú ateistickú odpoveď dáva Sam Harris: „Ty aj ja obaja vieme, že by nám trvalo päť minút vyrobiť knihu, ktorá ponúka súvislejšiu a súcitnejšiu morálku než Biblia.“ Podobne, človek môže skúsiť tvrdiť, že si váži Bibliu ako literárne dielo. Prečo si potom neváži *Pána Prsteňov*, čo je omnoho kvalitnejšie literárne dielo? A napriek štandardnej kritike Tolkienovej morálky je *Pán Prsteňov* oproti Biblii lepší prinajlepšom ako zdroj etiky. Prečo teda ľudia nenosia okolo krku prstienky namiesto krížikov? Dokonca aj *Harry Potter* je lepší ako Biblia, aj ako literárne dielo, aj ako morálna filozofia. Keby som chcel byť naozaj krutý, porovnal by som Bibliu so sériou *Kushiel* od Jacqueline Carey.

„Ako môžeš zdôvodniť, že si kupuješ laptop vykladaný drahokamami za 1 milión dolárov,“[→] opýtate sa priateľa, „keď mnoho ľudí nemá vôbec žiaden laptop?“ A váš priateľ povie: „Ale pomysli na pracovné miesta, ktoré to poskytne – výrobcovi laptopov, reklamnej agentúre výrobcu laptopov – a oni si potom kúpia jedlo a účesy – to bude stimulovať ekonomiku a nakoniec mnohí ľudia budú mať svoje vlastné laptopy.“ Bolo by však *efektívnejšie* kúpiť 5000 laptopov projektu One Laptop Per Child, čím by ste poskytli zamestnanie výrobcom OLPC *a zároveň* by ste priamo rozdali laptopy.

Už som sa dotkol témy zlyhania v hľadaní tretej možnosti. Ale toto nie je ozajstné motivované zastavenie. Nazvať to „motivované zastavenie“ by naznačovalo, že sa vôbec niekedy niečo hľadalo.

V kapitole Spodný riadok som ukázal, že iba skutočné príčiny našich rozhodnutí dokážu ovplyvniť našu efektivitu pri dosahovaní svojich cieľov. Niekto, kto si kupuje laptop za milión dolárov, si v skutočnosti myslel: „Aha, ligoce sa“ a to bola jediná kauzálna história jeho rozhodnutia kúpiť si laptop. Žiadne množstvo „zdôvodnenia“ to nezmení, pokiaľ to zdôvodnenie nie je skutočný proces nového vyhľadávania, ktoré by mohlo zmeniť záver. *Naozaj* zmeniť záver. Väčšina kritiky, ktorú robíme z pocitu povinnosti, je skôr symbolickou kontrolou než niečím iným. Ako slobodné voľby v krajine s jedinou politickou stranou.

Aby ste naozaj zdôvodnili úctyhodnosť Biblie jej literárnou kvalitou, museli by ste najprv nejako neutrálne prečítať kandidujúce knihy, dokiaľ nenájdete knihu s najvyššou literárnou kvalitou. Použijúc preslávenosť ako jedno možné rozumné kritérium na generovanie kandidátov navrhujem, aby ste si legitímne prečítali Shakespeara, Bibliu, a *Gödel, Escher, Bach*. (Inak by bola príliš veľká náhoda, aby sa Biblia vyskytla ako kandidát pri výbere z miliónov kníh.) Skutočne zložitá je tá časť „neutrálne prečítať“. Pomerne ľahké, ak nie ste kresťan, lenže ak ste...

Lenže nič takéto sa samozrejme nestalo. Nikto nič nehľadal. Napísať zdôvodnenie „literárna kvalita“ nad spodný riadok „milujem Bibliu“ je historické prekrútenie skutočnosti, ako sa tam ten spodný riadok naozaj dostal, je to ako predávať mačacie mlieko ako kravské mlieko. Toto nie je to, odkiaľ ten spodný riadok naozaj prišiel. Toto nie je to, čo sa pôvodne stalo a vytvorilo tento záver.

→ http://lesswrong.com/lw/km/motivated_stopping_and_motivated_continuation/

→ http://entertainment.slashdot.org/story/07/03/26/197253/a-million-dollar-laptop-created

Ak naozaj podrobíte svoj záver kritike, ktorá ho môže potenciálne od-záverovať – ak tá kritika *naozaj* má túto moc – to potom mení „skutočný algoritmus“ vášho záveru. To mení previazanosť vášho záveru v možných svetoch. Ale ľudia veľmi preceňujú *skutočnú* pravdepodobnosť, že zmenia svoj názor.

Pri toľkých otvorených mysliach by si človek myslel, že tu bude viac aktualizovania názorov.

Skúsim hádať: Áno, pripúšťate, že ste sa pôvodne rozhodli kúpiť si miliióndolárový laptop na základe myšlienky: „Aha, ligoce sa.“ Áno, pripúšťate, že toto nie je rozhodovací proces v súlade s vašimi vyjadrenými cieľmi. Lenže odvtedy ste sa rozhodli, že by ste v skutočnosti mali minúť svoje peniaze takým spôsobom, ktorý poskytne laptopy čo najväčšiemu množstvu chudákov bez laptopov. A predsa ste *nedokázali* nájsť efektívnejší spôsob, ako to dosiahnuť, než kúpiť si miliióndolárový laptop vykladaný diamantmi – pretože tak predsa dáte peniaze obchodu s laptopmi a stimulujete tak ekonomiku! To sa nedá prekonať!

Drahý priateľ, čertovsky podozrievam túto úžasnú zhodu okolností. Čertovsky podozrievam to, že najlepšia odpoveď podľa tohto milého, rozumného, altruistického kritéria X je zároveň tá istá myšlienka, aká vám pôvodne napadla na základe nesúvisiaceho, neobhájiteľného procesu Y. Ak si nemyslíte, že ste správnu otázku pravdepodobne získali hodom kockou, ako pravdepodobne by ste ju získali na základe iného nerozumného myslenia?

Je nepravdepodobné, že používate pomýlené uvažovanie, a predsa nerobíte žiadne chyby.

*→

—

77. Je toto vaše skutočné odmietnutie?

Z času na čas sa stane, že niekto natrafí na niektorý z mojich názorov ohľadom transhumanizmu – na rozdiel od mojich názorov týkajúcich sa ľudskej rozumnosti – čudnú, exoticky znejúcu myšlienku ako superinteligencia a Priateľská UI. A dotyčný ich odmietne.

Keď je potom vyzvaný, aby toto odmietnutie vysvetlil, neraz povie: „Prečo by som mal veriť niečomu z toho, čo Yudkowsky hovorí? Veď nemá PhD!“

A občas niekto iný, kto to počuje, povie: „Ach, mal by si si urobiť PhD, aby ťa ľudia počúvali.“ Alebo túto radu poskytne priamo ten, kto mi neverí, slovami: „Vráť sa až budeš mať PhD.“

Nuž, existujú dobré i zlé dôvody urobiť si PhD, ale toto je jeden z tých zlých.

Existuje veľa dôvodov, prečo niekto *naozaj* nepriateľsky reaguje na transhumanistické myšlienky. Väčšinou je to otázka rozoznávania vzorov, nie slovného uvažovania: myšlienka sa podobá na vzory „divná myšlienka“ alebo „vedecká fantastika“ alebo „sekta ohlasujúca koniec sveta“ alebo „prehnané mladícke nadšenie“.

Preto je okamžite, rýchlosťou vnemu, táto myšlienka odmietnutá. Ak sa potom niekto opýta: „Prečo nie?“, spustí sa tým hľadanie zdôvodnenia. Ale toto hľadanie nemusí nevyhnutne natrafiť na skutočný dôvod – pričom slovami „skutočný dôvod“ nemyslím *najlepší* dôvod, ktorý sa dá poskytnúť, ale skôr tú príčinu, ktorá ako historický fakt rozhodla v *prvom* momente, keď nastalo odmietnutie.

Namiesto toho hľadanie zdôvodnenia natrafí na fakt znejúci ako zdôvodnenie: „Tento rečník nemá PhD.“

Lenže ja nemám PhD ani vtedy, keď rozprávam o ľudskej rozumnosti, prečo teda aj tam neprichádza tá istá námietka?

A ešte viac k veci, *keby* som mal PhD, ľudia by to nepovažovali za rozhodujúci faktor naznačujúci, že by mali veriť všetkému, čo poviem. Namiesto toho by nastalo rovnaké počiatočné odmietnutie, z rovnakého dôvodu; a dodatočné hľadanie zdôvodnenia by sa potom zastavilo v odlišnom bode.

Povedali by: „Prečo by sme mali veriť *tebe*? Ty si len nejaký chlap s PhD! Takých je na svete kopa. Vráť sa, až sa preslávíš vo svojej oblasti a získaš definitívu na významnej univerzite.“

Ale veria ľudia *naozaj* náhodným profesorom z Harvardu, keď hovoria čudné veci? Samozrejme nie. (Keby som však bol profesorom na Harvarde, bolo by naozaj jednoduchšie získať *pozornosť médií*.

→ http://lesswrong.com/lw/kq/fake_justification/

Reportéri, ktorí by mali od začiatku sklon neveriť mi – ktorí by pravdepodobne mali rovnaký sklon neveriť náhodnému človeku s PhD – by aj tak o mne napísali článok, pretože aj to by bola novinka, že profesor z Harvardu verí takú divnú vec.)

Ak hovoríte veci, ktoré začiatočníkovi znejú *nesprávne*, na rozdiel od iba vychrlenia magicky znejúcich technoblábolov o splietaní leptických kvarkov v N+2 rozmeroch; a ak je poslucháč cudzinec, ktorý osobne nepozná *ani* vás *ani* hlavnú tému vášho odboru; potom si myslím, že bod, v ktorom náhodný človek *naozaj* začne prikladať dôveryhodnosť prekračujúcu jeho pôvodný dojem, iba *kvôli* akademickým dokladom, je niekde na úrovni nositeľa Nobelovej ceny. Ak vôbec. Približne povedané, potrebujete takú úroveň akademických dokladov, ktorá vás posúva „mimo dosah smrteľníkov".

Pokiaľ viem, približne toto sa stalo Ericovi Drexlerovi. Predložil svoju predstavu nanotechnológie a ľudia povedali: „Kde sú technické detaily?" a „Vráť sa, až budeš mať PhD!" A Eric Drexler strávil šesť rokov spisovaním technických detailov a dostal za to PhD pod vedením Marvina Minskeho. A *Nanosystémy* je skvelá kniha. Ale zmenili tí ľudia, ktorí hovorili: „Vráť sa, až budeš mať PhD", naozaj svoj názor na molekulárnu nanotechnológiu? Pokiaľ viem, nie.

Podobne bolo všeobecným pravidlom v Machine Intelligence Research Institute [Ústave výskumu strojovej inteligencie], že čokoľvek sa od nás žiadalo, aby sme boli dôveryhodnejší, keď sme to naozaj urobili, takmer nič sa nezmenilo. „Vyvíjate nejaký kód? Nemám záujem podporovať organizáciu, ktorá nevyvíja kód" -> OpenCog -> nič sa nezmenilo. „Eliezer Yudkowsky nemá akademické tituly" -> ako riaditeľ výskumu bol dosadený profesor Ben Goertzel -> nič sa nezmenilo. Jediná vec, ktorá naozaj *vyzerá*, že zvýšila dôveryhodnosť, sú známi ľudia spájaní s organizáciou, napríklad náš sponzor Peter Thiel, alebo člen rady Ray Kurzweil.

Toto môže byť dôležitá vec, ktorú sa oplatí zapamätať začínajúcim podnikateľov a čerstvým konzultantom – čo vám tí, ktorých ste si nezískali ako zákazníkov, *povedia* ako dôvod odmietnutia, nemusí *v skutočnosti* robiť žiaden rozdiel; a mali by ste sa nad tým starostlivo zamyslieť skôr než vynaložíte obrovské úsilie. Ak vám rizikový kapitalista povie: „Keby len váš predaj rástol o čosi rýchlejšie!", alebo ak vám potenciálny zákazník povie: „Vyzerá to dobre, ale nemá to vlastnosť X," nemusí to byť ich *skutočné* odmietnutie. Napraviť túto vec môže niečo zmeniť, ale nemusí.

A bolo by dobré pamätať na to počas hádok. Robin a ja máme spoločný názor, že dvaja racionalisti by sa nemali zhodnúť na tom, že sa nezhodnú: nemali by mať spoločné poznanie epistemického nesúhlasu, pokiaľ niečo nie je veľmi zle.

Predpokladám, že ak sa vo všeobecnosti dvaja racionalisti rozhodnú vyriešiť nezhodu, ktorá pretrvala po prvej výmene, mali by očakávať, že skutočné dôvody nesúhlasu sa buď ťažko komunikujú, alebo sa ťažko odhaľujú. Napríklad:

- Nezvyčajné, ale dobre podložené vedecké poznanie alebo matematika;
- Dlhé inferenčné vzdialenosti;
- Intuícia ťažko vyjadriteľná slovami, možno vychádzajúca z konkrétnych vizualizácií;
- Duch doby získaný zo svojej profesie (ktorá na to môže mať dobrý dôvod);
- Vnemovo rozoznané vzory na základe skúsenosti;
- Číre myšlienkové zvyky;
- Emocionálna pripútanosť veriť konkrétnemu výsledku;
- Strach z odhalenia minulých chýb;
- Sebaklam za účelom pýchy alebo iných osobných výhod.

Keby išlo o záležitosť, v ktorej sa *všetky* skutočné odmietnutia dajú *ľahko* vyložiť na stôl, nezhoda by sa pravdepodobne vyriešila tak priamočiaro, že by nezostala po prvom stretnutí.

„Je toto moje skutočné odmietnutie?" je otázka, ktorú by obaja nesúhlasiaci mali určite položiť *sami sebe*, aby to uľahčili tomu druhému. Avšak snahy priamo, verejne psychoanalyzovať toho druhého podľa môjho pozorovania zvyknú konverzáciu *veľmi* rýchlo zvrhnúť.

Aj tak by však malo by povolené, aby sa nesúhlasiaci pokorne opýtal: „Je toto tvoje skutočné odmietnutie?", ak existuje nejaký produktívny spôsob, ako sa venovať tejto podtéme. Možno by pravidlo malo znieť tak, že sa môžete otvorene spýtať: „Je tento jednoduchý priamočiaro znejúci dôvod tvoje

skutočné odmietnutie, alebo je to skôr intuícia X, alebo profesionálny duch doby Y?“ Zatiaľ čo tie zahanbujúcejšie možnosti uvedené naspodku tabuľky sú ponechané na svedomie toho druhého, a je jeho zodpovednosť, ako sa s tým vyrovná.

*→

78. Previazané pravdy, nákazlivé lži

> Jeden z našich veľmi raných filozofov došiel k záveru, že plne kompetentná myseľ by zo štúdia jedného faktu alebo predmetu patriaceho do nejakého vesmíru dokázala zostaviť alebo si predstaviť tento vesmír, od okamihu jeho stvorenia až po jeho úplný koniec...
>
> --First Lensman[89]

> Ak sa niekto z vás sústredí na jediný fakt alebo malý predmet, ako napríklad kamienok alebo semienko rastliny alebo iného tvora, na celkom krátke obdobie sto vašich rokov, začnete vnímať jeho pravdu.
>
> --Gray Lensman[90]

Som si pomerne istý, že jediný kamienok vzatý z pláže na našej vlastnej Zemi by neurčoval svetadiely a krajiny, politiku a ľudí na tento Zemi. Iné planéty v priestore a čase, iné Everettove vetvy, by vytvorili rovnaký kamienok. Na druhej strane, totožnosť jedného kamienka by zrejme zahŕňala naše fyzikálne zákony. V tom zmysle by celý náš vesmír – *všetky* Everettove vetvy – vyplývali z tohto kamienka. (Ak, ako sa zdá pravdepodobné, neexistujú žiadne naozaj voľné premenné.)

Jediný kamienok teda pravdepodobne neurčuje našu celú Zem. Ale z jediného kamienka vyplýva veľmi veľa. Zo štúdia toho jediného kamienka by ste mohli uvidieť zákony fyziky a všetko, čo vyplýva z nich. Rozmýšľajúc o týchto zákonoch fyziky by ste mohli uvidieť, ako sa tvoria planéty, a mohli by ste uhádnuť, že kamienok pochádza z niektorej takejto planéty. Vnútorné kryštály a molekulárne útvary kamienka vytvorené pod vplyvom gravitácie by vám povedali niečo o hmotnosti planéty; zmes prvkov v kamienku by vám povedala niečo o zložení planéty.

Nie som geológ, takže neviem, aké tajomstvá sú geológom k dispozícii. Ale viem si veľmi ľahko predstaviť, že ukážem geológovi kamienok a poviem: „Tento kamienok je z pláže na Half Moon Bay“, na čo mi geológ ihneď povie: „Som zmätený“ alebo priam „Klameš“. Možno je to nesprávny druh kameňa, alebo ten kamienok nie je dostatočne zodraný na to, aby mohol byť z pláže – ja nepoznám kamienky dosť dobre na to, aby som uhádol súvislosti a podpisy, podľa ktorých by ma mohli prichytiť, a práve to je pointa.

„Iba Boh dokáže povedať naozaj dôveryhodnú lož.“ Zaujímalo by ma, či niekedy existovalo náboženstvo, ktoré by si vytvorilo takéto príslovie. Tipol by som si (falzifikovateľne), že nie: je to racionalistický postoj, aj keď ho vyjadríte ako teologickú metaforu. Povedať: „všetko súvisí so všetkým, pretože Boh stvoril celý svet a udržiava ho“ môže vyvolať pekné hrejivé pocity počas kázne, ale nedostane vás to veľmi ďaleko, keď dôjde na priraďovanie kamienkov k plážam.

Minca na Zemi vytvorí gravitačné zrýchlenie na Mesiaci okolo $4{,}5 \times 10^{-31}$ m/s^2, takže v istom zmysle nie je celkom nesprávne povedať, že každá udalosť je previazaná s celým svojím svetelným kužeľom minulosti. A keďže odvodenia sa môžu prenášať kauzálnymi sieťami dopredu aj dozadu, *epistemické* previazania môžu ľahko prekročiť hranice svetelných kužeľov. Ale nechcem by som byť forenzným astronómom, ktorý sa musí pozrieť na Mesiac a zistiť, či na tej minci padla hlava alebo znak – ten vzťah je menší než kvantová neistota a tepelný šum.

Keby ste povedali: „Všetko je previazané s niečím iným“ alebo „Všetko je inferenčne previazané a niektoré previazania sú omnoho silnejšie než iné,“ môžete byť naozaj múdry a nie iba Hlboko Múdry.

→ http://lesswrong.com/lw/wj/is_that_your_true_rejection/

89 Edward Elmer Smith and A. J. Donnell, *First Lensman* (Old Earth Books, 1997).

90 Edward Elmer Smith and Ric Binkley, *Gray Lensman* (Old Earth Books, 1998).

Fyzicky je každá udalosť v istom zmysle súčtom celého svojho svetelného kužeľa minulosti, bez hraníc. Ale zoznam *pozorovateľných* previazaní je omnoho kratší a dáva vám niečo ako sieť. Táto pravidelnosť na vysokej úrovni je to, o čom hovorím, keď poviem Veľká Sieť Kauzality.

Používam tieto Veľké Písmená tak trochu ironicky; ale ak si niečo zaslúži Veľké Písmená, jednoznačne je to Veľká Sieť Kauzality.

„Ach, akú zamotanú sieť to tkáme, keď sa prvýkrát klamať pokúšame," povedal Sir Walter Scott. Nie *všetky* lži sa nám vymykajú spod kontroly – nežijeme v takom spravodlivom vesmíre. Ale občas sa stane, že niekto klame o nejakom fakte, a potom musí klamať o previazanom fakte, a potom o ďalšom fakte, ktorý je previazaný s týmto:

„Kde si bol?"

„Ehm, bol som na služobnej ceste."

„O čom bola tá služobná cesta?"

„To ti nemôžem povedať; to bolo dôverné jednanie s významným zákazníkom."

„Och – teba na také pozývajú? Dobrá správa! Mal by som zavolať tvojmu šéfovi a poďakovať mu za to, že ťa tam vzal."

„Prepáč – teraz práve nie je vo svojej kancelári..."

Ľuďom, ktorí nie sú bohovia, sa často nepodarí *predstaviť* si všetky fakty, ktoré by museli prekrútiť, aby povedali naozaj uveriteľnú lož. „Boh spôsobil, že som tehotná" znelo o trochu dôveryhodnejšie za starých čias, predtým než náš model sveta začal obsahovať (pojem) chromozóm Y. Mnohé podobné lži môžu dnes prasknúť, keď sa genetické testovanie stáva čoraz častejším. Ľudia boli odsúdení za znásilnenie, a falošné obvinenia boli odhalené, o niekoľko rokov neskôr, na základe indícií, o ktorých si neuvedomovali, že ich nechávajú. Študent evolučnej biológie môže vidieť rukopis prirodzeného výberu na každom vlkovi, ktorý naháňa zajaca; a na každom zajacovi, ktorý uteká preč; a na každej včele, ktorá uštipne namiesto toho, aby slušne upozornila – ale samotným kreacionistom *ich vlastné* lži znejú dôveryhodne, to som si istý.

Nie všetky lži sú odhalené, nie všetci klamári sú potrestaní; nežijeme v takom spravodlivom vesmíre. Ale nie všetky lži sú také bezpečné, ako klamári veria. Ktovie, koľko hriechov by odhalila bayesovská superinteligencia, keby urobila (nedeštruktívny?) nanotechnologický sken Zeme? Prinajmenšom všetky tie lži, o ktorých stále existujú indície v nejakom mozgu. Niektoré také lži sa môžu stať známymi ešte skôr, ak sa neurovedcom niekedy podarí zostaviť naozaj dobrý detektor lži pomocou neurozobrazovania. Paul Ekman (priekopník v štúdiu drobných pohybov tvárových svalov) by pravdepodobne teraz dokázal prečítať veľký zlomok svetových lží, keby dostal príležitosť.

Nie všetky lži sú odhalené, nie všetci klamári sú potrestaní. Ale Veľká Sieť zvykne byť často podceňovaná. Naštudovanie len tých vedomostí, ktoré ľudia *už zhromaždili*, by trvalo mnoho ľudských životov. Každý, kto si myslí, že nielen Boh dokáže povedať dokonalú lož, bez rizika, podceňuje zamotanosť Veľkej Siete.

Je úprimnosť najlepšou taktikou? Neviem, či chcem zájsť až tak ďaleko: Aj v mojej etike je niekedy okej držať hubu. Ale v porovnaní s priamymi lžami, aj úprimnosť aj mlčanie zahŕňa menšie vystavenie sa rekurzívnemu propagovaniu rizík, o ktorých neviete, že ich podstupujete.

*→

79. O klamstvách a výbuchoch čiernych labutí

Sudca Marcus Einfeld, 70-ročný, kráľovský advokát od roku 1977, austrálsky živý poklad 1997, cena mieru OSN 2002, zakladajúci predseda Austrálskej komisie pre ľudské práva a rovnosť príležitostí, ktorý pred niekoľkými rokmi odišiel na dôchodok, ale pravidelne je privolávaný súdiť dôležité prípady...

→ http://lesswrong.com/lw/uw/entangled_truths_contagious_lies/

...išiel do väzenia→ na dva roky za sériu krivých prísah a klamstiev, ktoré začali 36-librovou pokutou za prekročenie povolenej rýchlosti o 10 km/h.→

Celá tá *podozrivo cnostne znejúca* teória o tom, že poctiví ľudia nevedia dobre klamať, a že niekde zostanú previazané stopy, a že celá vec môže vybuchnúť v epickom zlyhaní typu čierna labuť, naozaj *má* niekoľko príkladov v skutočnom živote, hoci tu samozrejme pracuje selektívne spravodajstvo, keď o tomto počujeme.

*→

80. Epistemológia temnej strany

Ak raz zaklamete, pravda bude už naveky vaším nepriateľom.

Už som hovoril o tom, že pravdy sú previazané a lži sú nákazlivé. Ak zdvihnete kamienok z cesty a poviete geológovi, že ste ho našli na pláži – nuž, viete *vy*, čo taký geológ vie o kameňoch? Ja nie. Ale predstavujem si, že kamienok ošúchavaný vodou nebude vyzerať ako kvapka stuhnutej lávy z výbuchu sopky. Viete, odkiaľ ten kamienok na vašej ceste naozaj pochádza? Vo vesmíre, ktorý sa riadi zákonmi, veci nesú stopy svojho miesta; v tejto sieti lož nezapadá. (V skutočnosti mi geológ v komentároch povedal, že väčšina kamienkov na cestách pochádza *z pláží*, takže by nevedel určiť rozdiel medzi kamienkom z cesty a kamienkom z pláže, ale vedel by povedať rozdiel medzi kamienkom z hory a kamienkom z cesty/pláže. Názorný príklad...)

Čo jednej mysli znie ako nesúvisiaca pravda – ktorá by sa ľahko dala nahradiť uveriteľnou lžou – môže byť určené tuctom spojení pohľadom väčšieho poznania. Kreacionistovi môže myšlienka, že život utváral „inteligentný dizajn“ namiesto „prirodzeného výberu“ znieť ako názov športového tímu, ktorému fandí. Pre biológa, hodnoverne argumentovať, že nejaký organizmus bol inteligentne navrhnutý, by vyžadovalo klamať o takmer každej stránke tohto organizmu. Aby ste hodnoverne argumentovali, že „ľudia“ boli inteligentne nadizajnovaní, museli by ste klamať o dizajne ľudskej sietnice, o architektúre ľudského mozgu, o bielkovinách spojených slabými van der Waalsovými silami namiesto silných kovalentných väzieb...

Alebo by ste jednoducho klamali o evolučnej teórii, čo je cesta, ktorú si vyberie väčšina kreacionistov. Namiesto klamania o spojených uzloch v sieti, klamú o *všeobecných* zákonoch riadiacich tieto spojenia.

A aby potom *toto* celé zakryli, klamú o pravidlách vedy – napríklad čo znamená nazývať niečo „teória“, alebo čo to znamená, keď vedec povie, že si niečím nie je absolútne istý.

Takže prechádzajú od klamania o konkrétnych faktoch, cez klamanie o všeobecných pravidlách, ku klamaniu o pravidlách rozmýšľania. Aby ste klamali o tom, či sa ľudia vyvinuli, musíte klamať o evolúcii; a potom musíte klamať o pravidlách vedy, ktoré určujú naše chápanie evolúcie.

Ale ako inak? Rovnako ako by človek nezapadal do kolektívu *naozaj* inteligentne navrhnutých životných foriem, a museli by ste klamať o evolúcii, aby to vyzeralo inak; takisto samotné názory o kreacionizme nezapadajú do vedy – nenašli by ste ich v dobre usporiadanej mysli, rovnako ako by ste nenašli palmy rásť na ľadovci. A preto musíte prelomiť bariéry, ktoré tomu bránia.

Čo nás vedie k prípade sebaklamu.

Jedna lož, ktorú poviete *sami sebe*, môže znieť celkom prijateľne, keď neviete nič o pravidlách, ktorými sa riadia myšlienky, prípadne ani len, že také pravidlá *existujú*; a táto voľba vyzerá svojvoľne, ako keď si vyberáte chuť zmrzliny, izolovane ako kamienok na pobreží...

...ale potom niekto spochybní váš názor, pomocou pravidiel rozmýšľania, ktoré sa *on* naučil. Povie: „Kde sú tvoje indície?“

A vy poviete: „Čože? Načo by som potreboval indície?“

→ http://en.wikipedia.org/wiki/Marcus_Einfeld#Criminal_conviction

→ http://news.bbc.co.uk/2/hi/uk_news/magazine/7967982.stm

→ http://lesswrong.com/lw/9a/of_lies_and_black_swan_blowups/

On povie: „Vo všeobecnosti si názory vyžadujú indície.“

Tento argument je jasný vojak bojujúci za opačnú stranu, ktorého musíte poraziť. Preto poviete: „Nesúhlasím! Nie všetky názory vyžadujú indície. Konkrétne, názory o drakoch si nevyžadujú indície. Pokiaľ ide o drakov, môžeš veriť hocičomu, čo len chceš. Takže ja nepotrebujem indície, aby som veril, že v mojej garáži je drak.“

A on povie: „Čo? Nemôžeš len tak vynechať drakov. Pravidlo, že názory vyžadujú indície má svoj dôvod. Aby si nakreslil správnu mapu mesta, musíš ísť cez ulice a robiť na papieri čiary zodpovedajúce tomu, čo vidíš. To nie je svojvoľné nariadenie – ak sedíš vo svojej obývačke a kreslíš na papieri čiary náhodne, mapa bude nesprávna. S extrémne vysokou pravdepodobnosťou. To platí rovnako o mape draka ako hocičoho iného.“

Takže teraz už aj *toto* vysvetlenie, *prečo* názory vyžadujú indície, je *ďalší* nepriateľský vojak. Preto poviete: „Nesprávne s extrémne vysokou pravdepodobnosťou? Takže je tam stále šanca, však? Nemusím ti veriť, ak to nie je absolútne isté.“

Alebo možno začnete mať sami podozrenie, že „názory si vyžadujú indície“. Lenže to ohrozuje lož, ktorá je vám drahá; preto odmietnete svitanie vo svojom vnútri a zatlačíte slnko späť pod horizont.

Alebo ste už predtým počuli príslovie „názory si vyžadujú indície“ a znelo vám to dosť múdro, a verejne ste to schvaľovali. Ale nikdy vám celkom nedoplo, dokiaľ vás na to niekto iný neupozornil, že by sa toto príslovie mohlo *vzťahovať* aj na váš názor, že vo vašej garáži je drak. Tak sa rýchlo zamyslíte a poviete: „Ten drak je oddelené magistérium.“

Mať nepravdivé názory nie je dobré, ale nemusí to byť trvalo poškodzujúce – ak svoju chybu po odhalení opustíte. Nebezpečná vec je mať nepravdivý názor, o ktorom *veríte, že musí byť chránený ako viera* – viera vo vieru, či už je sprevádzaná skutočnou vierou alebo nie.

Jediná Lož, Ktorú Treba Chrániť dokáže zablokovať niekoho postup do pokročilej rozumnosti. Nie, nie je to neškodná zábava.

Tak ako je samotný svet omnoho previazanejší než sa na povrchu zdá; rovnako sú aj prísnejšie pravidlá rozmýšľania, obmedzujúce názory silnejšie, než by netrénovaný človek predpokladal. Svet je pevne prepletený, riadený všeobecnými zákonmi, a takisto aj *rozumné* názory.

Predstavte si, čo by to vyžadovalo poprieť evolúciu alebo heliocentrizmus – všetky tie previazané pravdy a riadiace zákony, ktoré by ste nemali dovolené poznať. Potom si dokážete predstaviť, ako jediný akt sebaklamu dokáže zablokovať celú meta-úroveň hľadania pravdy, keď sa vaša myseľ začne cítiť ohrozená videním súvislostí. Zakáže všetky stredne pokročilé a vyššie úrovne racionalistovho Umenia. Namiesto toho vytvorí rozľahlú sústavu antizákonov, pravidiel antimyslenia, všeobecné zdôvodnenia prečo veriť nepravdivému.

Steven Kaas povedal: „Šíriť menej než maximálne presné názory je akt sabotáže. Nerobte to nikomu, pokiaľ by ste mu zároveň nechceli prezerať pneumatiky.“ Dať niekomu *ochraňovať* nepravdivý názor – presvedčiť ho, že *samotný tento názor* treba chrániť pred každou myšlienkou, ktorá vyzerá, že ho ohrozuje – nuž, toto by ste nemali robiť nikomu, pokiaľ by ste mu zároveň nechceli urobiť frontálnu lobotómiu.

Keď raz zaklamete, pravda je vaším nepriateľom, aj každá pravda spojená s touto pravdou, a každý spojenec pravdy vo všeobecnosti; proti všetkému z tohto musíte bojovať, aby ste tú lož ochránili. Či už klamete druhým alebo sám sebe.

Musíte popierať, že názory si vyžadujú indície, a potom musíte popierať, že by mapy mali odrážať územie, a potom musíte popierať, že pravda je dobrá vec...

Takto vzniká Temná Strana.

Obávam sa, že ľudia si to neuvedomujú, alebo že nie sú dostatočne opatrní – že ako chodíme po našom ľudskom svete, môžeme očakávať, že sa stretneme so *systematicky* zlou epistemológiou.

Mémy „ako myslieť“ sa vznášajú okolo nás, uložené myšlienky Hlbokej Múdrosti – niektoré z nich sú dobré rady vymyslené racionalistami. Ale iné predstavy boli vymyslené, aby chránili lož alebo sebaklam: výtvory Temnej Strany.

„Každý má právo na svoj vlastný názor.“ Keď sa nad tým zamyslíte, kde vzniklo toto príslovie? Je to niečo, čo by niekto povedal v procese chránenia pravdy, alebo v procese chránenia *pred* pravdou? Ale ľudia nevyskočia a nepovedia: „Aha! Cítim tu prítomnosť Temnej Strany!“ Pokiaľ viem, väčšina si ani neuvedomuje, že Temná Strana existuje.

Ale ako inak? Či klamete druhých alebo iba seba, Lož, Ktorú Treba Chrániť sa bude rekurzívne šíriť po sieti empirickej kauzality, po sieti všeobecných empirických pravidiel, a pravidiel samotného rozmýšľania, a porozumení týchto pravidiel. Ak na svete existuje *dobrá* epistemológia, a zároveň aj lži alebo sebaklamy, ktoré sa ľudia snažia chrániť, potom vznikne aj zlá epistemológia, aby odporovala tej dobrej. Sotva môžeme v tomto svete očakávať, že nájdeme Svetlú Stranu bez Temnej Strany; existuje Slnko, aj to, čo sa od neho odťahuje a vytvára si ochranný Tieň.

Pripomína, že to nemusia nevyhnutne byť *zlí* ľudia. Prevažná väčšina z tých, ktorí opakujú Hlbokú Múdrosť sú viac oklamaní než neúprimní, viac sebaklamúci než klamúci. Aspoň si to myslím.

A určite nie je mojím zámerom poskytnúť vám Úplne Všeobecný Protiargument, aby ste hocikedy, keď vám niekto ponúkne epistemológiu, ktorá sa vám nepáči, mohli povedať: „Ach, toto vymyslel niekto z Temnej Strany.“ Je jedno z pravidiel Svetlej Strany, že návrh musíte odmietnuť kvôli nemu samotnému, nie obvinením jeho autora zo zlých úmyslov.

Lenže Temná strana existuje. Strach je tá cesta, ktorá k nej vedie, a jedna zrada vás môže obrátiť. Nie všetci, ktorí nosia rúcha, sú buď Jediovia alebo podvodníci; existujú aj Sithovia, majstri a nevedomí učni. Dávajte si pozor, buďte opatrní.

Pokiaľ ide o vymenovanie častých mémov, ktoré vytvorila Temná Strana – pripomínam, nie náhodné nepravdivé názory, ale zlá epistemológia, Všeobecné Obrany Zlyhania – nuž, chcelo by sa vám do toho pustiť, milí čitatelia?

*→

→ http://lesswrong.com/lw/uy/dark_side_epistemology/

H: Proti doublethinku

81. Singlethink

Pamätám si ten presný okamih, keď som začal svoje putovanie ako racionalista.

Nebolo to počas čítania *To nemyslíte vážne, pán Feynman*, ani žiadneho existujúceho diela o rozumnosti; tie som jednoducho prijal ako samozrejmé. Putovanie začne, keď uvidíte veľkú chybu vo svojom existujúcom umení a objavíte túžbu zlepšiť, vytvoriť *nové* zručnosti presahujúce tie užitočné ale nedostačujúce, ktoré ste našli v knihách.

V posledných okamihoch svojho prvého života som mal pätnásť rokov a opakoval som si príjemnú spomienku na vlastnú spravodlivosť z čias, keď som bol oveľa mladší. Moje spomienky na také dávne času sú hmlisté; mám v myšlienkach predstavu, ale nepamätám si, koľko rokov som presne mal. Myslím, že to bolo šesť alebo sedem, a že pôvodná udalosť sa stala počas letného tábora.

Pôvodne sa stalo to, že táborový vedúci, teenager, povedal nám, omnoho mladším chlapcom, aby sme sa zoradili do radu, a navrhol nasledujúcu hru: chlapec na konci radu sa bude plaziť pomedzi naše nohy, a my ho budeme plieskať po zadku, keď pôjde okolo, a potom bude na rade ďalší osemročný chlapec na konci radu. (Možno je to tým, že som stratil svoju detskú nevinnosť, ale nemôžem sa ubrániť úvahám...) Odmietol som hrať túto hru, tak som bol poslaný sedieť do kúta.

Táto spomienka – odmietania biť a byť bitý – mi začala symbolizovať, že dokonca v takomto veľmi ranom veku som odmietal mať radosť z ubližovania druhým. Že som nebol ochotný kúpiť si úder po cudzom zadku za cenu úderu po mojom vlastnom; nebol som ochotný platiť bolesťou za príležitosť spôsobiť bolesť. Odmietol som hrať hry so záporným súčtom.

A potom, vo veku pätnásť rokov, som si náhle uvedomil, že to nie je pravda. *Neodmietol* som kvôli principiálnemu postoju proti hrám s nulovým súčtom. O väzenskej dileme som sa dočítal pomerne skoro vo svojom živote, ale nie ako sedemročný. Odmietol som jednoducho preto, lebo som nechcel, aby mi ubližovali, a stáť v rohu bola prijateľná cena za to, že mi nebudú ubližovať.

Čo je dôležitejšie, uvedomil si, že som toto *vždy* vedel – že skutočná spomienka *vždy* číhala v kútiku mojej mysle, moje myšlienkové oko sa na ňu na zlomok sekundy zahľadelo a potom sa odvrátilo preč.

V mojom prvom kroku po Ceste som *prichytil tento pocit* – zovšeobecnený nad subjektívnu skúsenosť – a povedal som: „Takže *takýto* je to pocit, keď zatlačím neželanú pravdu do kúta svojej mysle! Odteraz si budem všímať vždy, keď to urobím, a vyčistím *všetky* svoje kúty!“

Toto cvičenie som si nazval *singlethink*, podľa Orwellovho doublethinku. V doublethinku zabudnete, a potom zabudnete, že ste zabudli. V singlethinku si všimnete, že zabúdate, a potom si spomeniete. V hlave naraz držíte iba jednu vnútorne neprotirečivú myšlienku.

„Singlethink“ bola prvá *nová* racionalistická zručnosť, ktorú som si vytvoril, o ktorej som predtým nečítal v knihách. Pochybujem, že je pôvodná v zmysle akademickej pôvodnosti, ale našťastie to nie je potrebné.

Och, moje pätnásťročné ja rado pomenovávalo veci.

Desivé hĺbky sklonu potvrdzovať idú ďalej a ďalej. Nie navždy, pretože váš mozog má konečnú zložitosť, ale dosť dlho na to, aby vám to pripadalo ako večnosť. Stále objavujete nové mechanizmy (alebo o nich čítate), ktorými váš mozog odstrkuje veci z cesty.

Ale moje mladé ja pomocou tejto prvej metly vymietlo pekných pár kútov.

*→

—

→ http://lesswrong.com/lw/k0/singlethink/

82. Doublethink (dobrovoľné skreslenie)

> Medzi O'Brienovými prstami sa zjavil podlhovastý ústrižok novín. Asi päť sekúnd bol vo Winstonovom zornom poli. Bola to fotografia, a nebolo pochýb o jej identite. Bola to tá fotografia. Bola to ďalšia kópia fotografie Jonesa, Aaronsona a Rutherforda na zasadnutí strany v New Yorku, ktorú pred jedenástimi rokmi náhodne zazrel a okamžite zničil. Iba krátku chvíľku bola pred jeho očami, potom bola opäť v nedohľadne. Ale videl ju, nepochybne ju videl! Urobil zúfalý bolestivý pokus oslobodiť svoju hornú polovicu tela. Nedalo sa pohnúť ani o centimeter žiadnym smerom. Na chvíľu dokonca zabudol aj na stupnicu. Všetko, čo chcel, bolo držať túto fotografiu ešte raz vo svojich prstoch, alebo ju aspoň vidieť.
>
> „Existuje!“ vykríkol.
>
> „Nie,“ povedal O'Brien.
>
> Prešiel krížom cez miestnosť.
>
> V protiľahlej stene bola pamäťová diera. O'Brien nadvihol mriežku. Útly prúžok papiera sa nevidený vznášal preč na prúde teplého vzduchu; zanikal v záblesku plameňa. O'Brien sa odvrátil od steny.
>
> „Popol,“ povedal. „Neidenfikovateľný popol. Prach. Neexistuje. Nikdy neexistovala.“
>
> „Ale existovala! Stále existuje! Existuje v pamäti. Pamätám si ju. Ty si ju pamätáš.“
>
> „Ja si ju nepamätám,“ povedal O'Brien.
>
> Winstonovo srdce pokleslo. Toto bol doublethink. Cítil sa na smrť bezmocný. Aj keby si bol istý, že O'Brien klame, pravdepodobne by na tom nezáležalo. Bolo však dokonale možné, že O'Brien na fotografiu naozaj zabudol. A ak áno, potom už zabudol aj na svoje popieranie, že si ju pamätá, a zabudol aj na samotný akt zabudnutia. Ako si človek mohol byť istý, že je to jednoducho trik? Možno sa myseľ naozaj mohla takto šialene vykĺbiť; to bola myšlienka, ktorá ho premohla.
>
> --George Orwell, 1984[91]

Čo ak nám sebaklam pomáha byť šťastnými? Čo ak sa rozbehneme a prekonáme skreslenia a urobí nás to – ach! – *nešťastnými*? Iste by pravou múdrosťou bola rozumnosť *druhého rádu*; vyberanie si, kedy byť rozumným. Potom by ste sa mohli rozhodnúť, ktoré kognitívne skreslenia vás budú ovládať, aby ste maximalizovali svoje šťastie.

Ponechajúc bokom morálku, pochybujem, že by sa myseľ naozaj mohla takto šialene vykĺbiť.

Z rozumnosti druhého rádu by vyplývalo, že si v istom bode pomyslíte: „A teraz budem nerozumne veriť, že vyhrám v lotérii, aby som sa urobil šťastným.“ My však nemáme takúto priamu kontrolu nad svojimi názormi. Nedokážete sa primäť veriť, že obloha je zelená, aktom vôle. Mohli by ste uveriť, že tomu veríte – hoci som vám to práve skomplikoval tým, že som poukázal na rozdiel. (Nie je za čo!) Mohli by ste dokonca *uveriť*, že ste šťastný a oklamaný; ale neboli by ste *naozaj* ani šťastný ani oklamaný.

Aby rozumnosť druhého rádu bola skutočne *rozumná*, najprv by ste potrebovali dobrý model skutočnosti, aby ste vedeli odvodiť dôsledky rozumnosti a nerozumnosti. Keby ste sa potom rozhodli byť v prvom ráde nerozumný, potrebovali by ste zabudnúť na tento presný pohľad. A potom zabudnúť na akt zabudnutia. Nechcem sa dopustiť logickej chyby zovšeobecňovania fiktívnej indície, ale myslím si, že Orwell celkom dobre extrapoloval, kam táto cesta vedie.

Nemôžete poznať dôsledky skreslenia, dokiaľ ste sa nezbavili skreslení. A potom už je príliš neskoro na sebaklam.

---

91 Orwell, *1984*.

Druhou možnosťou je rozhodnúť sa zostať slepo skreslený, bez akejkoľvek jasnej predstavy dôsledkov. To nie je rozumnosť druhého rádu. To je zámerná hlúposť.

Buďte nerozumne optimistickí ohľadom svojich vodičských schopností, a budete veselo bezstarostní tam, kde sa ostatní budú triasť od strachu. Nebudete sa ani zaťažovať nepohodlným bezpečnostným pásom. Budete veselo bezstarostní celý deň, týždeň, rok. Potom TRESK, a zvyšok života strávite želaním si, aby ste si mohli poškrabať svrbiacu fantómovú končatinu. Alebo paralyzovaní od krku nižšie. Alebo mŕtvi. Nie je to nevyhnutné, ale je to možné. Aké je to pravdepodobné? Takýto obchod nemôžete urobiť rozumne, dokiaľ nepoznáte svoje *skutočné* vodičské schopnosti, aby ste vedeli zistiť, akému veľkému nebezpečenstvu sa vystavujete. Takýto obchod nemôžete urobiť rozumne, dokiaľ neviete o skresleniach ako je zanedbanie pravdepodobnosti.

Nezáleží na tom, koľko dní strávite v blaženej nevedomosti, jediná chyba stačí na zničenie ľudského života, na vyváženie každého haliera, ktorý ste zodvihli z koľajníc hlúposti.

Jedna z hlavných rád, ktoré dávam ašpirujúcim racionalistom, je: „Nepokúšaj sa byť chytrý.“ A tiež: „Počúvaj tie tiché, otravujúce pochybnosti.“ Ak neviete, potom neviete, *čo* neviete; neviete, *koľko* toho neviete; a neviete, ako veľmi *potrebujete* vedieť.

Neexistuje rozumnosť druhého rádu. Existuje iba skok naslepo, ktorý môže alebo nemusí viesť do jamy s blčiacou lávou. Keď toto *viete*, už je neskoro na slepotu.

Ale ľudia to ignorujú, pretože nevedia, čo nevedia. Neznáme neznáme veci im nie sú dostupné. Nesústredia sa na prázdne miesto na mape, ale berú ho akoby zodpovedalo prázdnemu územiu. Keď sa rozhodnú naslepo skočiť, skontrolujú si, čo majú v pamäti nejaké nebezpečenstvo, ale na prázdnej mape nenájdu žiadnu jamu s blčiacou lávou. Prečo neskočiť?

Bol som tam. Skúšal som to. Popálil som sa. Nesnažte sa byť chytrý.

Jednej kamarátke som raz povedal, že mám podozrenie, že šťastie z hlúposti je veľmi preceňované. Ona vážne pokrútila hlavou a povedala: „Nie, nie je; naozaj nie je.“

Možno niekde existujú hlúpi šťastní ľudia. Možno sú šťastnejší než vy. Ale život nie je fér, a vy sa nestanete šťastnejšími, keď budete žiarliť na to, čo nemáte. Mám podozrenie, že prevažná väčšina čitateľov *Overcoming Bias* by nedokázala dosiahnuť „šťastie z hlúposti“, keby to skúšali. Táto cesta je vám uzavretá. Nikdy nedosiahnete tento stupeň nevedomosti, nemôžete zabudnúť to, čo viete, nemôžete nevidieť to, čo ste videli.

Šťastná hlúposť je pre vás uzavretá. Nikdy ju nedosiahnete, okrem skutočného poškodenia mozgu, a možno ani vtedy. Mohli by ste sa azda zamýšľať, či je šťastie z hlúposti *optimálne* – či to nie je najväčšie šťastie, o aké sa človek môže snažiť – ale na tom nezáleží. Pre vás je táto cesta uzavretá, ak vôbec niekedy otvorená bola.

Všetko, čo vám teraz zostáva, je snažiť sa o takú šťastie, aké môže dosiahnuť racionalista. Myslím si, že sa v konečnom dôsledku môže ukázať ako väčšie. Sú tam ohraničené cesty a cesty s otvoreným koncom; rovinky, na ktorých možno leňošiť, a hory, na ktoré treba vyliezť; a ak to lezenie zaberá veľa úsilia, o to vyššia je tá hora.

Navyše, k životu patrí viac než len šťastie; a od vašich rozhodnutí môže závisieť šťastie nielen vás samotných.

Ale to všetko je akademická otázka. V čase, keď si uvedomíte, že ste mali na výber, už na výber nemáte. Nedokážete nevidieť to, čo ste raz videli. Tá druhá cesta je uzavretá.

*→

83. Nie, naozaj, ja som sa oklamal

Nedávno som sa rozprával s osobou, ktorá... ťažko to opísať. Formálne bola ortodoxná židovka. Bola aj vysoko inteligentná, vedomá si niektorých archeologických indícií proti jej náboženstvu, a štandardných plytkých argumentov proti náboženstvu, o ktorých veriaci vedia. Napríklad vedela,

→ http://lesswrong.com/lw/je/doublethink_choosing_to_be_biased/

že Mordecai, Ester, Haman a Vašti neboli v perzských historických záznamoch, ale že existovala zodpovedajúca stará perzská legenda o babylonských bohoch Marduk a Ištar, a súperiacich elamitských bohoch Humman a Vašti. *Vedela* to, a predsa oslavovala Purim. Jeden z tých vysoko inteligentných veriacich ľudí, ktorí sa celé roky dusia vo svojich rozporoch, rozoberajú a prekrúcajú, dokiaľ ich myseľ nevyzerá ako vnútro kresby M. C. Eschera.

Väčšina takýchto ľudí bude predstierať, že sú príliš múdri na to, aby sa rozprávali s ateistami, ale ona bola ochotná so mnou pár hodín debatovať.

Vďaka tomu teraz rozumiem prinajmenšom jednu ďalšiu vec o sebaklame, ktorej som predtým explicitne nerozumel – konkrétne, že nie je potrebné, aby ste *naozaj* oklamali sami seba, dokiaľ *veríte*, že ste sami seba oklamali. Môžete to nazvať „viera v sebaklam“.

Keď bola táto žena na strednej škole, myslela si, že je ateistka. Ale vtedy sa rozhodla, že sa bude správať, akoby verila v Boha. A potom – povedala mi vážne – postupom času začala naozaj veriť v Boha.

Pokiaľ môžem povedať, v tomto bode sa úplne mýli. Počas nášho rozhovoru znovu a znovu hovorila: „*Verím* v Boha“, ale ani jediný raz nepovedala: „Boh *existuje*.“ Keď som sa jej pýtal, prečo je veriaca, ani raz nehovorila o dôsledkoch existencie Boha, iba o dôsledkoch viery v Boha. Nikdy: „Boh mi pomôže,“ ale vždy: „moja viera v Boha mi pomáha.“ Keď som povedal: „Keby niekomu záležalo iba na pravde, a pozrel by sa na náš vesmír, nebol by ani vymyslel Boha ako hypotézu,“ jednoznačne so mnou súhlasila.

Nepresvedčila *naozaj* samu seba, aby uverila, že Boh existuje alebo že židovské náboženstvo je pravdivé. Ani zďaleka nie, pokiaľ môžem povedať.

Na druhej strane si myslím, že *naozaj* verí, že sama seba oklamala.

Takže aj keď nemá žiaden úžitok z viery v Boha – pretože neverí – úprimne *verí*, že oklamala sama seba na vieru v Boha, a preto úprimne *očakáva*, že bude mať úžitok, ktorý so spája s oklamaním seba samej na vieru v Boha; a *toto* by asi malo tvoriť zhruba rovnaké placebo ako *naozajstná* viera v Boha.

A to by mohlo vysvetliť, prečo mala motiváciu zapálene brániť tvrdenie, že *verí* v Boha, pred mojím skeptickým spochybňovaním, aj keď nikdy nepovedala: „Ach, mimochodom, Boh naozaj existuje“, ani nevyzerala, že by sa čo len trochu o takýto výrok zaujímala.

*→

84. Viera v sebaklam

Písal som o svojom rozhovore s formálnou ortodoxnou židovkou, ktorá energicky obhajovala tvrdenie, že verí v Boha, pričom vyzerala, že v Boha v skutočnosti vôbec neverí.

Kým som sa jej pýtal, aké výhody podľa nej má viera v Boha, povedal som jej Tarskeho litániu – čo je vlastne nekonečná množina litánií, z ktorých konkrétny príklad je:

> Ak je obloha modrá,
>
> Chcem veriť: „Obloha je modrá.“
>
> Ak obloha nie je modrá,
>
> Chcem veriť: „Obloha nie je modrá.“

„Toto nie je moja filozofia,“ povedala mi.

„Nemyslel som si, že je,“ odpovedal som jej. „Len sa pýtam – za predpokladu, že by Boh *neexistoval*, a že by sa o tom vedelo, či by ste aj tak mali veriť v Boha?“

Zaváhala. Zdalo sa, že sa naozaj pokúša rozmýšľať o tom, čo ma prekvapilo.

„Takže je to hypotetická otázka...“ povedala pomaly.

Vtedy som si myslel, že má problém dovoliť si predstaviť si svet, v ktorom Boh neexistuje, pretože je pripútaná k svetu, v ktorom Boh existuje.

→ http://lesswrong.com/lw/r/no_really_ive_deceived_myself/

Teraz si skôr myslím, že mala problém predstaviť si kontrast medzi tým, ako by vyzeral *svet*, keby Boh existoval a keby neexistoval, pretože všetky jej myšlienky boli o jej *viere v Boha*, ale jej kauzálna sieť modelovala svet, ktorý Boha ako uzol neobsahoval. Dokázala by teda ľahko odpovedať na: „Ako by sa tento svet odlišoval, keby som neverila v Boha?“ ale nie: „Ako by sa tento svet odlišoval, keby Boh neexistoval?“

Vtedy mi na túto otázku neodpovedala. Ale našla *protipríklad* k Tarskeho litánii:

Povedala: „Verím, že ľudia sú lepší, než naozaj sú.“

Pokúsil som sa jej vysvetliť, že keď poviete: „Ľudia sú zlí,“ znamená to, že veríte, že ľudia sú zlí, a keď poviete: „Verím, že ľudia sú dobrí,“ znamená to, že veríte že veríte, že ľudia sú dobrí. Takže povedať: „Ľudia sú zlí a ja verím, že ľudia sú dobrí,“ znamená, že veríte, že ľudia sú zlí, ale veríte že veríte, že ľudia sú dobrí.

Zacitoval som jej:

> Keby existovalo sloveso s významom „veriť nepravdivo“, nemalo by v prítomnom čase prvú osobu jednotného čísla.
>
> --Ludwig Wittgenstein[92]

Ona povedala, s úsmevom: „Áno, ja verím, že ľudia sú lepší než naozaj sú. Len som ti to chcela takto vysvetliť.“

> „Myslím si, že by sa Babička mala na teba pozrieť, Walter,“ povedala Nanny. „Myslím si, že tvoja myseľ je pomotaná ako klbko nití, ktoré spadlo.“
>
> --Terry Pratchett, Maškaráda[93]

A ja dokážem napísať slová: „Nuž, myslím si, že neverila tomu, že by jej myslenie malo byť reflexívne konzistentné,“ ale stále mám problém sa s tým vyrovnať.

Dokážem vidieť vzorce v slovách, ktoré vychádzajú z jej úst, ale nedokážem na úrovni empatie pochopiť myseľ, ktorá je za nimi. Viem sa vžiť do kože mimozemšťanov, ktorí jedia deti a Tretej Lady Kiritsugu, ale nedokážem si predstaviť, aké to musí byť na jej mieste. Alebo iba *nechcem*?

Toto je dôvod, prečo inteligentní ľudia majú iba určité množstvo času (meraného v subjektívnom čase rozmýšľania o náboženstve), aby sa stali ateistami. Po istom bode, ak ste chytrý, ak ste strávili ten čas myslením na vaše náboženstvo a jeho obhajobu, a ak ste stále neunikli z pazúrov Epistemológie temnej strany, vaša myseľ zvnútra skončí ako Escherov obrázok.

(Jeden z ďalších pár momentov, keď sa zamyslela – spomínam to, keby ste mali príležitosť to použiť – bol, keď hovorila o tom, aké dobré je veriť, že sa niekto zaujíma o to, či robíte dobré alebo zlé veci – samozrejme *nie* o tom, že naozaj *existuje* Boh, ktorý sa zaujíma o tom, či robíte dobré alebo zlé veci, takýto výrok nebol súčasťou jej náboženstva...

A ja som povedal: „Ale mne záleží na tom, či robíte dobro alebo zlo. Takže vlastne hovoríte, že toto nestačí, a že potrebujete zároveň veriť v niečo *nad* ľudstvom, čomu záleží na tom, či robíte dobro alebo zlo.“ To ju na chvíľu zastavilo, pretože samozrejme týmto spôsobom o tom predtým nerozmýšľala. Iba štandardné použitie neštandardnej sady nástrojov.)

Neskôr som sa jej v istom bode opýtal, či by bolo dobré robiť *čokoľvek* odlišne, keby Boh jednoznačne neexistoval, a tentokrát odpovedala: „Nie.“

„Takže,“ povedal som neveriaco, „či Boh existuje alebo neexistuje, to nemá absolútne žiaden účinok na to, ako by ľudia mali rozmýšľať alebo konať? Myslím si, že aj rabín by sa na toto pozeral podozrievavo.“

Zdá sa, že jej náboženstvo sa teraz skladá *výlučne* z uctievania samotného uctievania. Kým skutoční veriaci starých čias asi verili, že ich zachráni všetko vidiaci otec, ona dnes verí, že ju zachráni viera v Boha.

---

92 Ludwig Wittgenstein, *Philosophical Investigations*, trans. Gertrude E. M. Anscombe (Oxford: Blackwell, 1953).

93 Terry Pratchett, *Maskerade*, Discworld Series (ISIS, 1997).

Potom, čo povedala: „Verím, že ľudia sú lepší než naozaj sú," som sa opýtal: „Takže ste pravidelne prekvapená, keď ľudia sklamú vaše očakávania?" Nasledovalo dlhé ticho, a potom, pomaly: „No... som *prekvapená*, keď ľudia... sklamú moje očakávania?"

Vtedy som tejto pauze nerozumel. Vtedy som tým chcel naznačiť, že ak je pravidelne sklamaná skutočnosťou, potom je toto nevýhoda viery v nepravdivú vec. Ale zdá sa, že ju zaskočili dôsledky toho, že *nie je* zaskočená.

Teraz si uvedomujem, že celá podstata jej filozofie bola *jej viera, že sama seba oklamala*, a možnosť, že jej odhady ľudí boli *v skutočnosti presné*, ohrozila Epistemológiu Temnej Strany, ktorú si postavila okolo názorov ako: „Mám úžitok z toho, že verím, že ľudia sú lepší než naozaj sú."

Zosadila z trónu starú modlu a nahradila ju vysloveným uctievaním Epistemológie Temnej Strany, ktorá pôvodne vznikla, aby túto modlu chránila; uctievala svoj vlastný pokus o sebaklam. Pokus zlyhal, ale ona si to úprimne neuvedomovala.

A tak musia symbolickí strážcovia príčetnosti ľudstva (motto: „kazíme vaše pomätené zábavky už od Epikurových čias") dnes bojovať proti aktívnemu uctievaniu sebaklamu – uctievaniu *domnelých výhod viery*, namiesto Boha.

Toto v skutočnosti vysvetľuje fakt o *mne*, ktorému som predtým nerozumel – dôvod, prečo ma vytáča, keď ľudia hovoria o tom, aký je sebaklam *ľahký*, a prečo píšem celé články na blogu o tom, že urobiť vedomé rozhodnutie veriť, že obloha je zelená, je omnoho ťažšie, než si ľudia asi myslia.

Je to preto, lebo – hoci sa *nemôžete* len tak rozhodnúť veriť, že obloha je zelená – ak si tento fakt *neuvedomíte*, v skutočnosti sa *môžete* oklamať, že veríte, že ste sa úspešne oklamali.

A keďže potom úprimne *očakávate*, že obdržíte výhody, o ktorých si myslíte, že pochádzajú zo sebaklamu, dostanete rovnaký druh placebo výhod, aké by ste boli naozaj dostali z úspešného sebaklamu.

Keď teda chodím a vysvetľujem, aký *ťažký* je sebaklam, v skutočnosti priamo mierim na tie placebo výhody, ktoré majú ľudia z viery, že sa úspešne oklamali, a útočím na nový druh náboženstva, ktoré uctieva iba uctievanie Boha.

Rozmýšľam, či táto bitka nevytvorí nový zoznam dôvodov, prečo nie viera, ale *samotná* viera vo vieru, je dobrá vec? Prečo ľudia majú veľký úžitok z uctievania svojho uctievania? Budeme to musieť robiť znovu s vierou vo vieru vo vieru a s uctievaním uctievania uctievania? Alebo sa inteligentní veriaci nakoniec skrátka vzdajú tejto línie argumentovania?

Rád by som veril, že nikto nemôže naozaj veriť, že verí, že verí, že verí, lenže filozofické argumenty o svete zombií sú už omnoho zamotanejšie než toto a ich zástancovia ich stále neopustili.

*→

85. Moorov paradox

Moorov paradox je štandardné označenie pre vetu: „Vonku prší, ale ja neverím, že prší."

Myslím si, že po prečítaní niektorých komentárov na *Less Wrong* rozumiem Moorovmu paradoxu o čosi lepšie. Jimrandomh hovorí:

> Mnoho ľudí nedokáže rozlíšiť medzi úrovňami nepriamosti. Pre nich „verím v X" a „X" sú to isté, a preto dôvody, prečo sa oplatí veriť v X, sú zároveň dôvodmi, prečo je X pravda.

Nemyslím si, že je to správne – relatívne malé deti dokážu pochopiť pojem, že majú nesprávny názor, čo si vyžaduje oddelené myšlienkové vedrá pre mapu a pre územie. Ale ukazuje to smerom k podobnej myšlienke:

Mnoho ľudí asi nedokáže vedome rozlíšiť medzi *vierou* v niečo a *schvaľovaním* niečoho.

Napokon... „verím v demokraciu" znamená, v bežnej reči, že schvaľujete tento systém, a nie že veríte, že demokracia existuje. Slovo „verím" má teda viac než jeden význam. Možno hľadáme

→ http://lesswrong.com/lw/s/belief_in_selfdeception/

zmätené slovo, ktoré spôsobuje zmätené myslenie (alebo možno len odráža predchádzajúce zmätené myslenie).

Takže: v pôvodnom príklade: „Verím že ľudia sú lepší než naozaj sú," uviedla nejaké dôvody, prečo by bolo dobré veriť, že ľudia sú dobrí – zdravotné výhody a podobne – a keďže potom mala hrejivý pocit ohľadom „veriť, že ľudia sú dobrí", preskúmala tento hrejivý pocit a došla k záveru: „Verím, že ľudia sú dobrí." Čiže si pomýlila *kladný pocit* priradený citovanému názoru za signál *jej viery v daný výrok*. Zároveň samotný svet vyzerá, akoby ľudia neboli takí dobrí. Preto povedala: „Verím, že ľudia sú lepší než naozaj sú."

A to už hraničí s poctivou chybou – v istom zmysle – keďže ľudia nie sú vyslovene učení, ako majú vedieť, že niečomu veria. Ako v podobenstve o drakovi v garáži; ten, kto hovorí: „V mojej garáži je drak – ale je neviditeľný", si neuvedomuje svoje *očakávanie*, že neuvidí žiadneho draka ako náznak, že má (správny) model bez draka.

Nie je to tak, že by ľudí *učili* rozoznať, kedy niečomu veria. Nie je to tak, že by ich niekedy učili na strednej škole: „Skutočne niečomu veriť – mať tento výrok vo svojom zozname presvedčení – vám pripadá tak, že to skrátka vyzerá, že presne takto svet *je*. Mali by ste rozoznať tento pocit, čo je skutočný (bez úvodzoviek) názor, a odlíšiť ho od toho, že máte dobrý pocit z nejakého názoru, ktorý rozoznávate ako názor (čo znamená, že je v úvodzovkách)."

Toto výrazne pomáha, aby nám príklady Moorovho paradoxu zo skutočného života pripadali menej cudzie, a poskytuje nám to ďalší mechanizmus, ktorým ľudia môžu zároveň mať aj nemať pravdu.

Podobne Kurige, ktorý napísal:

> Verím, že existuje Boh – a že nám vštepil zmysel pre dobro a zlo, ktorým dokážeme vyhodnocovať svet okolo nás. Zároveň verím, že zmysel pre morálku nám naprogramovala evolúcia – zmysel pre morálku, ktorý je s najväčšou pravdepodobnosťou výsledkom vytvárania zoskupení a meta-politických koalícií v komunitách bonobov veľmi, veľmi dávno. Tieto dva názory si neodporujú, ale je zložité ich oba zosúladiť.

Obávam sa, Kurige, že si sa rozhodol, že máš *dôvody podporovať* citovaný názor, že nám Boh vštepil zmysel pre dobro a zlo. A tiež, že máš dôvody schvaľovať verdikt vedy. Obe z toho vyzerajú byť dobré komunity, do ktorých sa chceš pridať, však? Obe tieto množiny názorov majú svoje výhody? Pozrieš sa do svojho vnútra a zistíš, že máš z oboch týchto názorov dobrý pocit?

Ale *nepovedal* si:

„Boh nám vštepil zmysel pre dobro a zlo, a zároveň nám evolúcia naprogramovala zmysel pre morálku. Tieto dva stavy skutočnosti si neodporujú, ale je zložité ich oba zosúladiť."

Ak toto čítaš, Kurige, mal by si veľmi rýchlo vysloviť uvedené nahlas, aby si si všimol, že to vyzerá aspoň o trochu ťažšie na prehltnutie – všimni si ten *subjektívny rozdiel* – skôr než si dáš tú námahu znovu racionalizovať.

Toto je subjektívny rozdiel medzi tým, keď máte dôvody schvaľovať dva odlišné názory, a keď máte mentálny model jedného sveta, jedného stavu, ako sa veci majú.

*→

86. Neverte, že sa sami oklamete

Nechcem vyzerať, že zapáram do Kurigeho, ale myslím si, že musíte očakávať isté množstvo spochybňovania, ak prídete na LessWrong a poviete:

> Jedna z vecí, ktoré som si uvedomil, ktoré pomáhajú vysvetliť rozdiel, ktorý cítim, keď hovorím s väčšinou ostatných kresťanov, je fakt, že niekde po ceste môj svetonázor urobil veľkú odbočku preč od slepej viery a skončil niekde v susedstve Orwellovského double-thinku.

→ http://lesswrong.com/lw/1f/moores_paradox/

„Ak *vieš*, že je to double-think...“

„...ako tomu ešte môžeš *veriť*?“ chce sa mi bezmocne povedať.

Alebo:

> Rozhodol som sa veriť v existenciu Boha – úmyselne a vedome. Toto rozhodnutie však má absolútne nulový vplyv na skutočnú existenciu Boha.

Ak *vieš*, že tvoj názor nekoreluje so skutočnosťou, ako ho ešte stále môžeš mať?

Nemalo by *inštinktívne* uvedomenie si: „Aha, moment, obloha naozaj *nie je* zelená,“ vyplývať z uvedomenia si: „Moja mapa, ktorá hovorí: ‚obloha je zelená‘ nemá žiaden dôvod korelovať s územím?“

Nuž... očividne nie.

Jedna časť tohto hlavolamu môže byť moje vysvetlenie Moorovho paradoxu („Prší, ale ja neverím, že prší“) - že si ľudia introspektívne mýlia pozitívny pocit spojený s citovaným názorom so skutočnou vierou.

Ale ďalšia časť toho môže skrátka byť, že – v rozpore s rozhorčením, ktoré som tu pôvodne chcel napísať – v skutočnosti je pomerne *ľahké* neurobiť skok z „Mapa, ktorá odráža územie by povedala: ‚X‘,“ k skutočnej viere v „X“. Vyžaduje si to istú prácu *vysvetliť* predstavu mysle ako konštruktéra súladu medzi mapou a územím, a ešte aj potom si môže vyžadovať veľa práce dostať tieto dôsledky na úroveň *inštinktu*.

Teraz si uvedomujem, že keď som napísal: „Nemôžete sa prinútiť veriť, že obloha je zelená, aktom vôle,“ nebol som nezaujatým spravodajcom o existujúcich faktoch. Snažil som sa vytvoriť sebanapĺňajúce proroctvo.

Môže byť múdre chodiť okolo a úmyselne opakovať: „Mne by double-think nefungoval! Hlboko vnútri by som vedel, že to nie je pravda! Ak viem, že moja mapa nemá dôvod byť korelovaná s územím, to znamená, že tomu neverím!“

Pretože vtedy – ak by vás to vôbec niekedy pokúšalo – vám rýchlo napadnú myšlienky: „Ale ja viem, že to nie je naozaj pravda!“ a „Nedokážem oklamať sám seba!“; a takto bude naozaj menšia pravdepodobnosť, že úspešne oklamete sami seba. Máte väčšiu šancu pochopiť, na inštinktívnej úrovni, že povedať sám sebe X ešte z X neurobí pravdu: a preto v skutočnosti nie-X.

Ak si budete stále hovoriť, že si *nemôžete* len tak zvoliť, že budete veriť, obloha je zelená – potom je menšia pravdepodobnosť, že sa úspešne oklamete na nejakej úrovni, či už v zmysle, že by ste tomu naozaj uverili, alebo že by ste upadli do Moorovho paradoxu, viery vo vieru, alebo viery v sebaklam.

Ak si budete stále hovoriť, že hlboko vnútri viete...

Ak si budete stále hovoriť, že by ste jednoducho pozreli na svoju zložito vypracovanú nepravdivú mapu a skrátka by ste vedeli, že je to nepravdivá mapa bez očakávanej korelácie s územím, a preto by ste, napriek jej zložitej konštrukcii nedokázali do nej investovať svoju dôverčivosť...

Ak si budete stále hovoriť, že reflexívna konzistencia prevládne a spôsobí, že prestanete veriť na základnej úrovni, akonáhle si na meta-úrovni uvedomíte, že mapa neodráža...

Keď naozaj príde na lámanie chleba... možno to naozaj nedokážete.

Keď príde na úmyselný sebaklam, musíte *veriť vo svoju vlastnú neschopnosť!*

Povedzte sami sebe, že toto úsilie je odsúdené na neúspech... *a ono bude!*

Je toto sila pozitívneho myslenia, alebo sila negatívneho myslenia? Tak či onak, vyzerá to ako rozumná opatrnosť.

* →

—

→ http://lesswrong.com/lw/1o/dont_believe_youll_selfdeceive/

I: Videnie čerstvými očami

87. Ukotvenie a prispôsobenie

Predstavte si, že pred vaším zrakom zatočím kolesom šťastia a vyjde číslo 65. Potom sa opýtam: Myslíte si, že percento afrických krajín, ktoré sú v OSN, je väčšie alebo menšie než toto číslo? Aké percento afrických krajín je podľa vás v OSN? Zamyslite sa nad týmito otázkami na chvíľu sami, ak chcete, ale prosím nepoužívajte Google.

Ďalej, skúste odhadnúť, počas 5 *sekúnd*, hodnotu nasledujúceho matematického výrazu. 5 sekúnd. Pripravení? Pozor... *Štart!*

1 × 2 × 3 × 4 × 5 × 6 × 7 × 8

Tversky a Kahneman zapisovali odhady pokusných osôb, ktoré videli na kolese šťastia rôzne čísla.[94] Mediánový odhad osôb, ktoré videli na kolese číslo 65 bol 45%; mediánový odhad osôb, ktoré videli číslo 10 bol 25%.

Súčasná teória pre tento a podobné pokusy je, že pokusné osoby prijmú počiatočné neinformatívne číslo ako svoj štartový bod alebo *kotvu*; a potom sa *prispôsobia* smerom nahor alebo nadol od svojho počiatočného odhadu, kým nedosiahnu odpoveď, ktorá „znie prijateľne“; a potom sa prestanú prispôsobovať. Typickým výsledkom je nedostatočné prispôsobenie od kotvy – vzdialenejšie čísla by tiež mohli byť „prijateľné“, ale človek sa zastaví pri prvej prijateľne znejúcej odpovedi.

Podobne, študenti, ktorým ukážeme „1 × 2 × 3 × 4 × 5 × 6 × 7 × 8“ dávajú mediánový odhad 512, zatiaľ čo študenti, ktorým ukážeme „8 × 7 × 6 × 5 × 4 × 3 × 2 × 1“ dávajú mediánový odhad 2 250. Motivujúca hypotéza bola, že študenti skúsia vynásobiť (alebo odhadom skombinovať) prvých pár činiteľov a potom výsledok prispôsobia nahor. V oboch prípadoch bolo prispôsobenie nedostatočné, v porovnaní so skutočnou hodnotou 40 320, ale prvá množina odhadov bola omnoho nedostatočnejšia, pretože začínali z nižšej kotvy.

Tversky a Kahneman hlásia, že ponúknutie finančnej odmeny za presnosť nezmenšilo efekt ukotvenia.

Strack and Mussweiler (1997) sa pýtali na rok, kedy Einstein prvýkrát navštívil Spojené Štáty.[95] Úplne neuveriteľné kotvy ako 1215 alebo 1992 spôsobili rovnako veľký efekt ukotvenia ako prijateľnejšie kotvy ako 1905 alebo 1939.

Toto má zrejmé použitie napríklad pri dohadovaní o plate, alebo nákupe auta. Neradím vám, aby ste to zneužívali, ale aby ste si dali pozor na tých, čo chcú zneužiť vás.

A tiež: Všímajte si svoje vlastné myslenie a snažte sa všimnúť si, kedy *prispôsobujete* číslo pri hľadaní odhadu.

Pokusy o zbavenie sa skreslenia ukotvením sa vo všeobecnosti ukázali ako neúčinné. Odporučil by som tieto dva: Po prvé, ak vám pôvodný odhad pripadá neprijateľný, skúste ho úplne zahodiť a prísť s novým odhadom, a nie posúvať sa od kotvy. Ale toto samotné nemusí byť dostatočné – pokusné osoby, ktoré dostali pokyn vyhýbať sa ukotveniu, to naďalej robili.[96] Takže po druhé, aj keď ste vyskúšali tú prvú metódu, skúste ešte pomyslieť na kotvu v opačnom smere – kotvu, ktorá je očividne príliš malá alebo príliš veľká (ak tá prvá bola príliš veľká alebo príliš malá) – a chvíľku na nej zotrvajte.

⁂→

94 Amos Tversky and Daniel Kahneman, „Judgment Under Uncertainty: Heuristics and Biases,“ [Usudzovanie pri neistote: Heuristiky a skreslenia] *Science* 185, no. 4157 (1974): 1124–1131, doi:10.1126/science.185.4157.1124.

95 Fritz Strack and Thomas Mussweiler, „Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility,“ [Vysvetlenie záhadného efektu ukotvenia: Mechanizmus výberovej dostupnosti] *Journal of Personality and Social Psychology* 73, no. 3 (1997): 437–446.

96 George A. Quattrone et al., „Explorations in Anchoring: The Effects of Prior Range, Anchor Extremity, and Suggestive Hints“ [Skúmania pri ukotvovaní: Efekt prvotného rozsahu, krajnej kotvy, a sugestívnych náznakov] (Unpublished manuscript, Stanford University, 1981).

→ http://lesswrong.com/lw/j7/anchoring_and_adjustment/

88. Priming a kontaminácia

Predstavte si, že požiadate pokusné osoby, aby stlačili jedno tlačidlo, ak postupnosť písmen tvorí slovo a druhé tlačidlo, ak písmená netvoria slovo. (Napr. „banack“ verzus „banán“.) Potom im ukážete postupnosť „voda“. Neskôr budú rýchlejšie identifikovať postupnosť „piť“ ako slovo. Toto je známe ako „kognitívny priming“; táto konkrétna forma by bola „sémantický priming“ alebo „pojmový priming“.

Fascinujúca vec na primingu je, že sa odohráva na takej nízkej úrovni – zvyšuje rýchlosť *rozoznania písmen ako tvoriacich slovo*, čo sa zrejme stane *skôr* než sa zamyslíte nad významom daného slova.

Priming odhaľuje aj masívnu paralelizáciu šírenia aktivácie: ak videnie „voda“ aktivuje slovo „piť“, pravdepodobne aktivuje aj „rieka“ alebo „pohár“ alebo „špliechať“... a táto aktivácia sa šíri cez významové súvislosti pojmov až naspäť k rozoznávaniu postupností písmen.

Priming je podvedomý a nezastaviteľný, artefakt ľudskej nervovej architektúry. Pokúšať sa zastaviť svoj vlastný priming je ako pokúšať sa zastaviť šírenie aktivácie svojich vlastných nervových obvodov. Skúste nahlas povedať farbu – nie význam, ale farbu – nasledujúcej postupnosti písmen: „ZELENÁ“ (napísané červeným písmom).

Mussweiler a Strack (2000) sa pýtali pokusných osôb nasledujúcu ukotvujúcu otázku: „Je priemerná ročná teplota v Nemecku vyššia alebo nižšia ako 5 stupňov Celzia / 20 stupňov Celzia?“[97] Neskôr, pri úlohe identifikovania slov, pokusné osoby, ktoré dostali kotvu 5 stupňov rýchlejšie identifikovali slová ako „zima“ a „sneh“, kým pokusné osoby s vysokou kotvou rýchlejšie identifikovali „teplo“ a „slnko“. Toto ukazuje mechanizmus neprispôsobivosti pri ukotvení: priming kompatibilných myšlienok a spomienok.

Všeobecnejší záver je, že *úplne neinformujúce, vedome nepravdivé* alebo *úplne nesúvisiace* „informácie“ môžu ovplyvniť odhady a rozhodnutia. V oblasti heuristík a skreslení je tento všeobecnejší jav známy ako *kontaminácia*.[98]

Raný výskum heuristík a skreslení objavil efekty ukotvenia, napríklad pokusné osoby dávajú nižšie (vyššie) odhady percenta krajín OSN nachádzajúci sa v Afrike podľa toho, či sa ich predtým pýtali, či je toto percento vyššie alebo nižšie ako 10 (65). Tento efekt sa pôvodne pripisoval tomu, že sa pokusné osoby posúvajú od kotvy ako počiatočného bodu, a zastavia sa akonáhle dosiahnu prijateľný interval; a posúvajú sa nedostatočne, pretože sa zastavia na jednom konci intervalu spoľahlivosti.[99]

Raná hypotéza Tverskeho a Kahnemana sa stále zdá byť v niektorých prípadoch správnym vysvetlením, obzvlášť keď pokusné osoby sami vytvoria úvodný odhad.[100] Moderný výskum však naznačuje, že väčšina ukotvenia je v skutočnosti kvôli kontaminácii, nie prispôsobovaniu sa posúvaním.

Váš obchod s potravinami asi má otravné nápisy ako „Maximum 12 kusov na zákazníka“ alebo „5 kusov za 10 dolárov“. Sú tieto nápisy efektívne pri ovplyvňovaní zákazníkov, aby kupovali väčšie množstvá? Asi si myslíte, že vás to neovplyvňuje. Ale *niekoho* to musí ovplyvňovať, pretože tieto nápisy dokázateľne fungujú, čo je dôvod, prečo ich obchody stále vyvesujú.[101]

Najhroznejšou stránkou kontaminácie je, že slúži ako ďalšia z tisícich tvárí sklonu potvrdzovať. Akonáhle sa nejaká myšlienka dostane do vašej hlavy, primuje informácie, ktoré sú s ňou kompatibilné –

---

97 Thomas Mussweiler and Fritz Strack, „Comparing Is Believing: A Selective Accessibility Model of Judgmental Anchoring,“ *European Review of Social Psychology* 10 (1 1999): 135–167, doi:10.1080/14792779943000044.

98 Gretchen B. Chapman and Eric J. Johnson, „Incorporating the Irrelevant: Anchors in Judgments of Belief and Value,“ [Zahŕňanie nepodstatného: Kotvy pri hodnotení názorov a hodnoty] in Gilovich, Griffin, and Kahneman, *Heuristics and Biases*, 120–138.

99 Tversky and Kahneman, „Judgment Under Uncertainty.“

100 Nicholas Epley and Thomas Gilovich, „Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self-Generated and Experimentor-Provided Anchors,“ [Vrátenie prispôsobenia do heuristiky ukotvenia a prispôsobenia: Diferenciálne spracovanie sebou vytvorených a experimentátorom poskytnutých kotiev] *Psychological Science* 12 (5 2001): 391–396, doi:10.1111/1467-9280.00372.

101 Brian Wansink, Robert J. Kent, and Stephen J. Hoch, „An Anchoring and Adjustment Model of Purchase Quantity Decisions,“ [Model ukotvenia a prispôsobenia pri rozhodnovaní o množstve nákupov] *Journal of Marketing Research* 35, no. 1 (1998): 71–81, http://www.jstor.org/stable/3151931.

a tým si zabezpečuje svoju trvalú existenciu. Nielenže máme selekčný tlak na vyhrávania politických hádok; sklon potvrdzovať je zabudovaný priamo do nášho hardwaru, asociatívne siete primujú kompatibilné myšlienky a spomienky. Nešťastný vedľajší účinok našej existencie ako tvorov s nervovou sústavou.

Jediný letmý obrázok môže stačiť na to, aby primoval rozoznávanie súvisiacich slov. Nemyslite si, že treba viac na to, aby sa sklon potvrdzovať rozbehol. Chce to iba ten jeden rýchly záblesk a spodný riadok je už určený, pretože svoje názory meníme omnoho zriedkavejšie než si myslíme...

*→

89. Veríme všetkému, čo nám povedia?

Niektoré rané experimenty o ukotvovaní a prispôsobovaní testovali, či *rozptyľovanie* pokusných osôb – ich kognitívne „vyťaženie“ požiadavkou, aby sledovali, či sa v postupnostiach číslic vyskytuje „5“ alebo také čosi – zníži prispôsobovanie, a tým zvýši vplyv kotvy. Väčšina experimentov asi potvrdzuje myšlienku, že kognitívna vyťaženosť zvyšuje ukotvovanie a kontamináciu vo všeobecnosti.

Všímajúc si hromadiace sa experimentálne výsledky – viac a viac zistení o kontaminácii zvýraznenej kognitívnou vyťaženosťou – Daniel Gilbert uvidel, ako sa vynára naozaj šialený vzor: Veríme *všetkému*, čo nám povedia?

Človek by si mohol prirodzene myslieť, že keď nám niekto povie nejaký výrok, najprv *pochopíme*, čo ten výrok znamená, potom sa nad tým výrokom *zamyslíme*, a nakoniec ho buď *prijmeme* alebo *odmietneme*. Tento napohľad samozrejmý model toku kognitívnych procesov sa tiahne už od Descarta. Ale Descartov súper, Spinoza, nesúhlasil. Spinoza naznačil, že najprv *pasívne prijmeme výrok v procese jeho pochopenia*, a až dodatočne *aktívne prestaneme veriť* výrokom, ktoré po zvážení odmietneme.

Posledných pár storočí filozofi viacmenej držali s Descartom, keďže jeho pohľad vyzeral viac, však viete, logicky a intuitívne. Ale Gilbert videl spôsob, ako otestovať Descartovu a Spinozovu hypotézu experimentálne.

Ak má pravdu Descartes, potom by rozptyľovanie pokusných osôb mali prekážať aj prijímaniu pravdivých tvrdení aj odmietaniu nepravdivých. Ak má pravdu Spinoza, potom by rozptyľovanie pokusných osôb malo spôsobiť, že si nepravdivé tvrdenia budú pamätať ako pravdivé, ale nemalo by spôsobiť, že si pravdivé tvrdenia budú pamätať ako nepravdivé.

Gilbert, Krull, a Malone (1990) potvrdili tento výsledok, ukazujúc, že keď pokusným osobám predkladali nové tvrdenia označené PRAVDA a NEPRAVDA, rozptyľovanie nemalo vplyv na identifikovanie pravdivých tvrdení (55% úspech bez vyrušovania, 58% s vyrušovaním), ale ovplyvňovalo identifikovanie nepravdivých tvrdení (55% bez vyrušovania, 35% s vyrušovaním).[102]

Ešte dramatickejšiu ilustráciu vyprodukoval v nasledujúcom experimente Gilbert, Tafarodi a Malone (1993).[103] Pokusné osoby nahlas čítali správy o zločine rolujúce sa na obrazovke, v ktorých farba textu označovala, či je konkrétne tvrdenie pravdivé alebo *nepravdivé*. Niektoré správy obsahovali *nepravdivé* tvrdenia, ktoré zvyšovali vážnosť zločinu, iné správy obsahovali *nepravdivé* tvrdenia, ktoré zmierňovali (ospravedlňovali) zločin. Pokusné osoby mali počas čítania správ o zločine zároveň sledovať postupnosti písmen a hľadať číslicu „5“ – to bola rozptyľujúca úloha vytvárajúca kognitívne vyťaženie. Nakoniec mali pokusné osoby odporučiť dĺžku väzenia pre daného zločinca, od 0 do 20 rokov.

Pokusné osoby v kognitívne vyťaženom stave odporučili v priemere 11,15 rokov väzenia pre zločincov s „vážnejším“ zločinom, čiže zločincov, ktorých správy obsahovali *výroky zvyšujúce vážnosť zločinu označené ako nepravdivé*. Vyťažené osoby odporučili v priemere 5,83 rokov väzenia pre

→ http://lesswrong.com/lw/k3/priming_and_contamination/

102 Daniel T. Gilbert, Douglas S. Krull, and Patrick S. Malone, „Unbelieving the Unbelievable: Some Problems in the Rejection of False Information,“ *Journal of Personality and Social Psychology* 59 (4 1990): 601–613, doi:10.1037/0022-3514.59.4.601.

103 Gilbert, Tafarodi, and Malone, „You Can’t Not Believe Everything You Read.“

zločincov, ktorých správy obsahovali *výroky ospravedlňujúce zločin označené ako nepravdivé*. Tento takmer dvojnásobný rozdiel bol, ako asi očakávate, štatisticky významný.

Nevyťažení účastníci čítali presne tie isté správy, s rovnakými *označeniami*, a okolo nich občas prešli rovnaké postupnosti číslic, ale oni nemali za úlohu sledovať číslicu „5“. Preto mohli venovať viac pozornosti „nevereniu“ výrokov *označených ako nepravdivé*. Títo nevyťažení účastníci odporúčali 7,03 rokov verzus 6,03 rokov pre zločincov, ktorých správy *nepravdivo zvyšovali vážnosť* alebo *nepravdivo ospravedlňovali*.

Článok Gilberta, Tafarodiho a Maloneho mal názov: „Nedokážete neveriť všetkému, čo čítate.“

To naznačuje – prinajmenšom – že by sme mali byť omnoho opatrnejší, keď sa vystavujeme nespoľahlivým informáciám, najmä ak popri tom robíme ešte niečo iné. Buďte opatrní, keď zazriete noviny v supermarkete.

*→

90. Uložené myšlienky

Jednou z najväčších záhad ľudského mozgu je, ako tá prekliata vec *vôbec dokáže* fungovať, keď väčšina neurónov vysiela signál 10-20-krát za sekundu, maximálne 200 Hz. V neurológii existuje „pravidlo sto krokov“, že každá predpokladaná operácia musí skončiť po *nanajvýš* 100 postupných krokoch – môžete byť paralelní, ako len chcete, ale nemôžete predpokladať viac ako 100 (a radšej menej) neurónových signálov po sebe.

Viete si predstaviť musieť programovať pomocou 100Hz procesorov, bez ohľadu na to, ako veľa by ste ich mali? Potrebovali by ste stovky miliárd procesorov, aby ste *vôbec niečo* urobili v reálnom čase.

Keby ste museli písať programy pracujúce v reálnom čase pre sto miliárd 100Hz procesorov, jeden trik, ktorý by ste používali, kde by sa len dalo, by bolo skladovanie výsledkov. To znamená, že by ste si odložili výsledky predchádzajúcich operácií a nabudúce by ste ich vyhľadali namiesto prepočítavania od nuly. A to je typický *nervový* prístup – rozoznávanie, asociácie, dopĺňanie vzoru.

Je rozumné predpokladať, že *väčšina* ľudského poznávania pozostáva z hľadania uložených údajov.

Táto myšlienka mi z času na čas prichádza na rozum.

Existuje krásny ilustrujúci príbeh, ktorý som si myslel, že som si odložil medzi záložky, ale neviem ho opäť nájsť: bol to príbeh o človeku, ktorého sused „vševed“ raz mimochodom tvrdil, že najlepší spôsob, ako odstrániť z domu komín je vybúrať krb, počkať, kým tehly spadnú o úroveň nižšie, vybúrať tieto tehly, a opakovať, kým nezmizne celý komín. O niekoľko rokov neskôr, keď tento muž chcel odstrániť svoj komín, táto uložená myšlienka číhala, pripravená zaútočiť...

A ako si ten muž dodatočne uvedomil – asi uhádnete, že to nedopadlo dobre – jeho sused nebol v týchto veciach nejako zvlášť znalý, ani dôveryhodný zdroj. Keby nad touto myšlienkou *zapochyboval*, pravdepodobne by si bol uvedomil, že je hlúpa. Niektoré uložené údaje je lepšie opäť prepočítať. Ale mozog dopĺňa vzor automaticky – a ak si vedome neuvedomíte, že daný vzor treba upraviť, zostane vám doplnený vzor.

Predpokladám, že keby táto myšlienka napadla tomu mužovi samotnému – keby sám *osobne* dostal tento skvostný nápad ako odstrániť komín – bol by sa nad touto myšlienkou zamyslel kritickejšie. Ale keď už niekto *iný* túto myšlienku celú premyslel, môžete ušetriť výpočtovú silu tým, že si uložíte ich *záver*, nie?

Špeciálne v modernej civilizácii nikto nedokáže myslieť dosť rýchlo na to, aby si premyslel svoje vlastné myšlienky. Keby ma ako dieťa pohodili v lese, kde by ma vychovali vlci alebo tichí roboti, sotva by ste ma poznali ako človeka. Nikto nedokáže myslieť dosť rýchlo ani na to, aby za jeden život zrekapituloval múdrosť kmeňa lovcov a zberačov, začínajúc od nuly. A čo sa týka múdrosti gramotnej civilizácie, zabudnite na to.

→ http://lesswrong.com/lw/k4/do_we_believe_everything_were_told/

Avšak odvrátenou stranou tohto celého je, že pravidelne vidím ľudí, ktorí sa pokúšajú o kritické myslenie, ako papagájujú uložené myšlienky, ktoré neboli vymyslené kritickými mysliteľmi.

Dobrým príkladom je skeptik, ktorý pripúšťa: „No, faktickým indíciami nemožno náboženstvo ani dokázať ani vyvrátiť.“ Ako som inde zdôraznil, toto je jednoducho nepravdivá teória pravdepodobnosti. A takisto je to jednoducho nepravda z pohľadu reálnej psychológie náboženstva – pred pár stáročiami by vás za takýto výrok upálili na hranici. Matka, ktorej dcéra má rakovinu, sa modlí: „Bože, prosím uzdrav moju dcéru,“ a nie: „Milý Bože, viem, že náboženstvo nesmie mať žiadne falzifikovateľné dôsledky, čo znamená, že nie je možné, aby si uzdravil moju dcéru, takže... no, v podstate sa modlím, aby som sa cítila lepšie, namiesto robenia niečoho, čo by mojej dcére mohlo naozaj pomôcť.“

Ale ľudia si prečítajú: „Faktickým indíciami nemožno náboženstvo ani dokázať ani vyvrátiť,“ a potom, keď nabudúce uvidia nejaký kus indície vyvracajúcej náboženstvo, ich mozog si doplní tento vzor. Dokonca aj niektorí ateisti opakujú túto absurditu bez zaváhania. Keby sa nad touto myšlienkou zamysleli sami, namiesto počutia od niekoho iného, boli by skeptickejší.

Smrť: doplňte vzor: „Smrť dáva životu zmysel.“

Je frustrujúce hovoriť s dobrými a slušnými ľuďmi – ľuďmi, ktorí by ani za tisíc rokov nikdy *spontánne* nepomysleli na vyhladenie ľudskej rasy – spomenúť tému existenciálneho rizika, a počuť, ako hovoria: „No, možno si ľudská rasa nezaslúži prežiť.“ Za tisíc rokov by ich nikdy napadlo zastreliť svoje vlastné dieťa, ktoré je súčasťou ľudskej rasy, ale mozog doplní tento vzor.

Aké vzory, ktoré ste si nikdy nevybrali, sa dopĺňajú vnútri vašej mysle?

Rozumnosť: doplňte vzor: „Láska nie je rozumná.“

Keby táto myšlienka niekedy napadla vám osobne, ako celkom nová myšlienka, ako by ste ju kriticky skúmali? Viem, čo by som povedal *ja*, ale čo by ste povedali *vy*? Môže byť ťažké to vidieť čerstvými očami. Pokúste sa zabrániť svojej mysli v doplnení vzoru štandardným, neprekvapujúcim, už známym spôsobom. Možno že neexistuje lepšia odpoveď než tá štandardná, ale nemôžete o tej odpovedi *rozmýšľať*, dokiaľ nezabránite svojmu mozgu v automatickom dopĺňaní odpovede.

Teraz, keď ste si toto prečítali, až bude nabudúce niekoho počuť, ako bez zaváhania opakuje mém, ktorý považujete za hlúpy alebo nepravdivý, pomyslíte si: „Uložené myšlienky.“ Môj názor je teraz vo vašej mysli, čaká na doplnenie vzoru. Ale je to pravda? Nedovoľte svojej mysli doplniť tento vzor! *Myslite!*

*→

91. Koľaje „mimo vychodených koľají“

Kedykoľvek vás niekto nabáda, aby ste „mysleli mimo vychodených koľají“, zvyčajne vám, *pre vaše pohodlie*, ukáže, kde presne sa „mimo vychodených koľají“ nachádza. Sranda, ako sa všetci nezávisláci obliekajú rovnako...

V Umelej Inteligencii má každý mimo odboru uložený výsledok *skvelej novej revolučnej myšlienky UI* – neurónové siete, ktoré fungujú rovnako ako ľudský mozog! Nová myšlienka UI: doplňte vzor: „Logickým UI sa napriek všetkých veľkým sľubom nepodarilo poskytnúť skutočnú inteligenciu celé desaťročia – potrebujeme neurónové siete!“

Táto uložená myšlienka tu je už asi tri desaťročia. Stále žiadna umelá inteligencia. Napriek tomu si akosi každý mimo odboru myslí, že neurónové siete sú Nová Myšlienka Ktorá Rúca Dominantnú Paradigmu, odkedy v 70-tych rokoch vynašli spätnú propagáciu. Hovorte mi o vašich starnúcich hipíkoch.

Imidž nezávisláka svojou podstatou nedovoľuje žiadne odchýlky od normy. Ak nechodíte v čiernom, ako majú ľudia vedieť, že ste trpiaci umelec? Ako majú ľudia rozoznať jedinečnosť, ak nezapadáte do štandardného vzoru ako má jedinečnosť vyzerať? Ako má hocikto rozoznať, že máte novú revolučnú predstavu o UI, ak sa netýka neurónových sietí?

→ http://lesswrong.com/lw/k5/cached_thoughts/

Ďalším príkladom rovnakého žánru je „podvratná“ literatúra, ktorá všetka znie rovnako, a stojí za ňou malá vzdorovitá liga rebelov, ktorí ovládajú celú katedru angličtiny. Ako sa Anonymous opýtal na blogu Scotta Aaronsona:

> Už niekedy nejaká podvratná literatúra, ktorú ste čítali, spôsobila, že ste zmenili niektorý zo svojich politických názorov?

Alebo ako si všimol Lizard:

> Revolúcia už bola v televízii. Revolúcia už má svoj *merchandise*. Revolúcia je spotrebný tovar, životný štýl v balíčku dostupný vo vašom miestnom hypermarkete. Za 19,95 dolára si kúpite čiernu masku, konzervu so sprejom, transparent „Smrť fašizmu“ a konto na vlastný blog, na ktorom môžete písať o policajnej brutalite, ktorú ste utrpeli, keď ste sa priviazali reťazou k požiarnemu hydrantu. Kapitalizmus sa naučil predávať antikapitalizmus.

Mnohí v Silicon Valley si všimli, že drvivá väčšina špekulatívnych kapitalistov v ľubovoľnej dobe naháňa rovnaké Revolučné Inovácie, a sú to tie Revolučné Inovácie, ktoré mali verejnú ponuku účastín pred šiestimi mesiacmi. To je *zvlášť* zdrvujúce pozorovanie špekulatívneho kapitálu, pretože tam je priamy ekonomický motív nenasledovať dav – buď niekto iný vyvíja rovnaký produkt, alebo niekto iný ponúka príliš veľa za startup. Steve Jurvetson mi raz povedal, že vo firme Draper Fisher Jurvetson súhlas dvoch partnerov stačí na nákup ľubovoľného startupu do 1,5 milióna dolárov. Ale ak sa *všetci* partneri zhodnú, že niečo vyzerá ako dobrý nápad, tak to neurobia. Kiež by grantové komisie mali toľko rozumu.

Problémom originality je, že aby ste ju dosiahli, musíte naozaj *myslieť*, a nenechať svoj mozog dopĺňať vzory. Neexistuje žiadna pohodlná značka „mimo vychodených koľají“, ku ktorej sa môžete ihneď rozbehnúť. Funguje to takmer ako zen – nemôžete naučiť satori pomocou slov, pretože satori je zážitok, keď slová zlyhávajú. Čím viac sa snažíte nasledovať slovné pokyny zenového majstra, tým ďalej ste od dosiahnutia prázdnej mysle.

Myslím, že existuje dôvod, prečo ľudia nedosiahnu novosť tým, že sa o ňu snažia. Vlastnosti ako pravdivosť alebo dobrý dizajn nezávisia od novosti: 2 + 2 = 4, naozaj, fakt, aj keď si to myslí každý. Ľudia, ktorí sa snažia objaviť pravdu alebo vymyslieť dobrý dizajn, môžu postupom času dosiahnuť tvorivosť. Nie každá zmena je vylepšenie, ale každé vylepšenie je zmena.

Každé vylepšenie je zmena, ale nie každá zmena je vylepšenie. Ten, kto hovorí: „Chcem postaviť originálnu pascu na myši!“ a nie: „Chcem postaviť optimálnu pascu na myši!“ si takmer vždy želá byť *vnímaný* ako originálny. „Originalita“ v tomto zmysle je svojou podstatou sociálna, pretože sa dá určiť iba porovnaním s druhými ľuďmi. Takže ich mozog jednoducho doplní štandardný vzor toho, čo sa vníma ako „originálne“ a ich priatelia súhlasne prikývnu a povedia, že je to podvratné.

Knihy o biznise vám vždy povedia, pre vaše pohodlie, kam presne sa podel váš syr. V opačnom prípade by bol čitateľ zanechaný opakujúc: „Kde je to ‚mimo vychodených koľají‘, kam mám ísť?“

Skutočné myslenie, podobne ako satori, je úkon mysle bez slov.

Poprední filozofi z Monty Python to vyjadrili najlepšie zo všetkých vo filme *Život Briana*:[104]

„Musíte myslieť sami za seba! Všetci ste jednotlivci!“

„Áno, všetci sme jednotlivci!“

„Všetci ste rôzni!“

„Áno, všetci sme rôzni!“

„Musíte si to vyriešiť sami!“

„Áno, musíme si to vyriešiť sami!“

⁂→

104 Graham Chapman et al., *Monty Python's The Life of Brian (of Nazareth)* (Eyre Methuen, 1979).

→ http://lesswrong.com/lw/k6/the_outside_the_box_box/

92. Originálny pohľad

Keď to Robert Pirsig vyjadril takto dobre, ja už len skopírujem, čo povedal. Neviem, či sa tento príbeh zakladá na skutočnosti alebo nie, ale v oboch prípadoch je pravdivý.[105]

> Mal problém so žiakmi, ktorí nemali čo povedať. Najprv si myslel, že je to lenivosť, ale neskôr sa ukázalo, že nie. Jednoducho nedokázali vymyslieť, čo povedať.
>
> Jedna z nich, dievča s okuliarmi s hrubými sklami, chcela napísať esej dĺžky päťsto slov o Spojených Štátoch. Bol zvyknutý na sklamania, ktoré nasledujú po takýchto výrokoch, tak jej navrhol, bez podceňovania, aby to zúžila iba na Bozeman.
>
> Keď prišiel termín práce, nemala ju a bola veľmi rozrušená. Skúšala a skúšala, ale nenapadlo jej nič, čo by mohla povedať.
>
> To ho zarazilo. Teraz *on* nevedel vymyslieť, čo povedať. Nastalo ticho a potom nečakaná odpoveď: „Zúž to iba na *hlavnú ulicu* v Bozemane.“ Bol to záblesk osvietenia.
>
> Usilovne prikývla a odišla. Ale hneď pred ďalšou hodinou sa vrátila so *skutočnou* úzkosťou, tentokrát so slzami, tá úzkosť tam asi bola už dlho. Stále jej nenapadalo nič, čo by sa dalo povedať, a nechápala prečo, ak nedokázala vymyslieť nič o *celom* Bozemane, by mala byť schopná vymyslieť niečo o jednej jeho ulici.
>
> Rozčúlil sa: „Ty sa *nepozeráš!*“ Spomenul si na svoje vlastné vyhodenie z univerzity, pretože toho povedal *príliš* veľa. Ku každému faktu existuje *nekonečne* veľa hypotéz. Čím viac sa *pozeráš*, tým viac ich *vidíš*. Ona sa naozaj nepozerala, a ešte stále to nejako nepochopila.
>
> Povedal jej nahnevane: „Zúž to na *prednú* časť *jednej* budovy na hlavnej ulici v Bozemane. Dom opery. Začni od ľavej hornej tehly.“
>
> Jej oči za hrubými sklami okuliarov sa naširoko otvorili.
>
> Na ďalšiu hodinu prišla so začudovaným pohľadom a odovzdala mu esej dĺžky päťsto slov o prednej časti Domu opery na hlavnej ulici v Bozemane, v štáte Montana. „Sedela som v stánku s hamburgermi na opačnej strane cesty,“ povedala, „a začala som písať o prvej tehle, potom o druhej tehle, a pri tretej tehle to zrazu všetko na mňa prišlo a nevedela som sa zastaviť. Mysleli si, že mi šibe, robili si zo mňa srandu, ale tu to je. Nerozumiem tomu.“
>
> To ani on, ale počas dlhých prechádzok po uliciach mesta o tom rozmýšľal a došiel k záveru, že ju očividne blokovala podobná prekážka aká paralyzovala jeho počas jeho prvého dňa učenia. Bola zablokovaná, pretože sa pokúšala písomne zopakovať veci, ktoré už počula, tak ako sa on počas prvého dňa pokúšal zopakovať veci, ktoré sa už rozhodol povedať. Nedokázala vymyslieť nič, čo by sa dalo napísať o Bozemane, pretože si nedokázala spomenúť na nič počuté, čo by stálo za zopakovanie. Zvláštne si neuvedomovala, že by sa mohla pozrieť a vidieť čerstvo sama za seba, a potom písať, bez sledovania v prvom rade, čo už bolo povedané. Zúženie na jednu tehlu zničilo tento blok, pretože bolo príliš jasné, že *musí* urobiť originálny a priamy pohľad.
>
> --Robert M. Pirsig, Zen a údržba motocykla

* →

93. Čudnejšie než história

Predstavte si, že by som vám povedal, že viem *naisto*, že nasledujúce tvrdenia sú pravdivé:

105 Pirsig, *Zen and the Art of Motorcycle Maintenance*.

→ http://lesswrong.com/lw/k7/original_seeing/

- Keď sa natriete istým *presne* daným odtieňom farby medzi modrou a zelenou, otočí sa sila gravitácie a budete padať smerom nahor.
- V budúcnosti bude obloha zaplnená miliardami vznášajúcich sa čiernych gúľ. Každá guľa bude väčšia než všetky vzducholode, ktoré kedy existovali, dokopy. Keď guli ponúknete peniaze, spustí sa z oblohy prostitút na bungee lane.
- Vaše vnúčatá si budú myslieť, že by bola nielen hlúposť, ale *zlo*, zatvárať zlodejov za mreže namiesto toho, aby dostali po zadku.

Pomysleli by ste si, že mi šibe, však?

Teraz si predstavte, že ste v roku 1901 a máte si vybrať, či budete viac veriť horeuvedeným alebo nasledujúcim tvrdeniam:

- Existuje absolútna hranica rýchlosti akou sa dva predmety môžu voči sebe pohybovať, a je to presne 670 616 629,2 míle za hodinu. Ak vyskočíte na vlak, ktorý ide takmer takto rýchlo a vystrelíte z okna, základné jednotky dĺžky sa zmenia, takže *vám* sa bude zdať, že guľka letí rýchlo pred vami, ale iní ľudia uvidia niečo iné. Aha, čas okolo vás sa tiež zmení.
- V budúcnosti bude existovať superprepojená globálna sieť miliónov sčítacích zariadení, každé z nich bude silnejšie než všetky sčítacie stroje pred rokom 1901 dohromady. Jedným z hlavných využití tejto siete bude prenos pohyblivých obrázkov lesbického sexu, predstierajúc, že sa skladajú z čísel.
- Vaše vnúčatá si budú myslieť, že by bola nielen hlúposť, ale *zlo*, povedať, že niekto nemôže byť prezidentom Spojených Štátov, pretože je čierny.

Inšpirované Robinovým komentárom: *„Zaujímalo by ma, či by niekto vedel dostatočne podrobne opísať fiktívny príbeh o alternatívnej skutočnosti, ktorú by naši predkovia nedokázali odlíšiť od tej skutočnej, aby sa ozrejmilo, aká prekvapujúca sa ukázala skutočnosť.“*

*→

94. Logická chyba zovšeobecňovania fiktívnej indície

Keď sa pokúšam o úvod do problematiky vyspelej UI, čo je prvá vec, ktorú počujem vo viac než polovici prípadov?

„Aha, myslíš ako vo filme *Terminátor* / *Matrix* / Asimovovi roboti!“

Odpovedám: „No, nie presne tak. Snažím sa vyhnúť logickej chybe zovšeobecňovania fiktívnej indície.“

Niektorí to pochopia hneď a zasmejú sa. Iní bránia svoje použitie daného príkladu a nesúhlasia, že je to chyba.

Čo je zlé na používaní filmov alebo románov ako východiskových bodov diskusie? Nikto predsa netvrdí, že je to *pravda*. Kde je tam lož, kde je racionalistický hriech? Science fiction predstavuje autorov pokus o vizualizáciu budúcnosti; prečo nevyužiť myslenie, ktoré už za nás urobil niekto iný, namiesto začínania od nuly?

Nie každý chybný krok v presnom tanci rozumnosti spočíva vo vyslovenej viere v nepravdu; existujú aj jemnejšie spôsoby, ako sa mýliť.

Po prvé, zbavme sa predstavy, že science fiction predstavuje plnohodnotný rozumný pokus o predpovedanie budúcnosti. Ešte aj tí najusilovnejší spisovatelia science fiction sú v prvom rade rozprávačmi; požiadavky na príbeh nie sú rovnaké ako požiadavky na predpoveď. Ako poukázal Nick Bostrom:[106]

> Kedy ste naposledy videli film o tom, ako ľudstvo zrazu vyhynulo (bez varovania, a bez nahradenia nejakou inou civilizáciou)? Aj keby takýto scenár bol omnoho

→ http://lesswrong.com/lw/j1/stranger_than_history/

106 Nick Bostrom, „Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards,“ *Journal of Evolution and Technology* 9 (2002), http://www.jetpress.org/volume9/risks.html.

pravdepodobnejší než scenár, kde ľudskí hrdinovia úspešne odrazia inváziu príšer alebo robotických vojakov, nebola by to veľká zábava sledovať.

Vo fikcii teda existujú konkrétne skreslenia. Ale pokúšať sa opraviť tieto konkrétne skreslenia nestačí. Príbeh *nikdy* nie je rozumným pokusom o analýzu, ani u tých najusilovnejších spisovateľov science fiction, pretože príbehy nepoužívajú pravdepodobnostné distribúcie. Ako ilustráciu uvediem:

> Bob Merkelthud sa obozretne prešmykol cez dvere mimozemskej kozmickej lode, obzerajúc sa doprava a potom doľava (alebo doľava a potom doprava), či tam zostala niektorá z obávaných Vesmírnych Príšer. Po jeho boku bola jediná zbraň, ktorá proti Vesmírnym Príšerám účinkovala: s pravdepodobnosťou 30% Vesmírny Meč ukutý z čistého titanu, s pravdepodobnosťou 20% obyčajný železný sochor, s pravdepodobnosťou 45% lesklý čierny disk nájdený v dymiacich ruinách Stonehenge, pravdepodobnosť 5% je rozdelená na priveľa drobných možností, ktoré tu nebudeme menovať.
>
> Merkelthud (aj keď je značná šanca, že tam namiesto neho bola Susan Wifflefoofer) urobil dva kroky dopredu alebo jeden krok dozadu, keď ohromný rev preťal ticho čiernej prechodovej komory! Alebo tiché bzučanie v bielej prechodovej komore. Hoci Amfer a Woofi (1997) tvrdia, že v tomto bode bol Merkelthud zožratý, Spacklebackle (2003) upozorňuje na to, že...

Postavy môžu byť nevedomé, ale samotný *autor* nemôže povedať čarovné slovo: „neviem“. Hlavný hrdina musí prejsť budúcnosťou po jednej čiare, plnej podrobností, ktoré dávajú príbehu telo, od Wifflefooferovej primerane futuristických postojov voči feminizmu, až po farbu jej náušníc.

Potom sú všetky tieto priťažujúce podrobnosti a pochybné predpoklady zabalené a označené krátkym názvom, vytvárajúc ilúziu jedného balíka.

Pri problémoch s veľkým priestorom odpovedí, najťažšie nie je *overiť* správnosť odpovede, ale jednoducho ju vôbec v priestore odpovedí nájsť. Ak niekto začne otázkou, či nás UI uloží alebo neuloží do puzdier ako vo filme „Matrix“, skáče tým na 100-bitový predpoklad, bez zodpovedajúcich 98 bitov indície potrebných na nájdenie ho v priestore odpovedí s možnosťou hodnou explicitného zváženia. Po získaní prvých 98 bitov indície by už stačilo len o pár bitov viac, aby sme túto možnosť povýšili na takmer istotu, čo vám hovorí niečo o čom, kde sa robí väčšina danej práce.

Tento „úvodný“ krok nájdenia možností hodných explicitného zváženia zahŕňa kroky ako: Brať do úvahy čo vieš a čo nevieš, čo vieš a čo nevieš predpovedať, vynakladať vedomé úsilie vyhnúť sa skresleniu absurdity a rozšíriť intervaly istoty, zvážiť ktoré otázky sú tie dôležité, pokúšať sa zohľadniť možné Čierne Labute a myslieť na (dovtedajšie) neznáme neznáme veci. Skočiť na „*Matrix:* áno alebo nie?“ toto všetko preskakuje.

Každý profesionálny vyjednávač vie, že ovládať pojmy diskusie je veľmi blízko k ovládaniu výsledku diskusie. Ak začnete myslením na *Matrix*, budete myslieť na pochodujúce armády robotov, ktoré porazia ľudí po dlhom boji... a nie na superinteligenciu, ktorá luskne nanotechnologickými prstami. Sústredíte sa na súboj „my proti nim“, nasmerujete pozornosť na otázky ako: „Kto vyhrá?“ a „Kto by mal vyhrať?“ a „Bude UI naozaj taká?“ Vznikne všeobecná atmosféra zábavy: „Aká je tvoja úžasná predstava budúcnosti?“

A v prázdnote sa stratia: úvahy o viacerých možných dizajnoch mysle, ktoré by „umelá inteligencia“ mohla implementovať; závislosť budúcnosti od jej počiatočných podmienok; sila nadľudskej inteligencie a argument pre jej nepredvídateľnosť; ľudia, ktorí majú seriózny pohľad na celú vec a snažia sa s tým niečo urobiť.

Keby sa nejaký zákerný kazič diskusií rozhodol, že je *preňho* najlepšie, keď bude nútiť diskutérov, aby začali vyvracaním *Terminátora*, bol by v takomto posunutí debaty úspešný. Pri debate o kontrole zbraní rečník NRA nechce byť uvedený ako „fanatický strieľač“, ani oponent zbraní nechce byť uvedený ako „advokát odzbrojenia obetí“. Prečo by ste vy mali dovoliť rovnaké skreslenie rámca od hollywoodskych scenáristov, hoci len náhodne?

Novinári mi nehovoria: „Budúcnosť bude ako film *2001*.“ Ale pýtajú sa: „Bude budúcnosť ako film *2001*, alebo ako film *U.I.*?“ To je asi rovnaký problém rámca ako keby sa pýtali: „Mali by sme okresať dávky pre postihnutých veteránov, alebo zvýšiť dane bohatým?“

V pravekom prostredí neexistovali kiná; čo ste videli na vlastné oči, to bola pravda. Krátky pohľad na jedno slovo nás dokáže primovať a urobiť kompatibilné myšlienky dostupnejšie, s dokázateľným silným vplyvom na odhady pravdepodobnosti. Koľko zmätku si myslíte, že vo vašom úsudku napácha dvojhodinový film? Bude dosť ťažké odstrániť tieto škody pomocou úmyselného sústredenia – načo si teda pozývať upíra do domu? V šachu alebo v go, každý premárnený ťah je strata; v rozumnosti, každý vplyv, ktorý nie je indícia, je (v priemere) entropia.

Darí sa filmovým divákom neveriť tomu, čo vidia? Pokiaľ viem, málo filmových divákov sa správa ako keby *priamo* pozorovali budúcnosť Zeme. Ľudia, ktorí videli filmy *Terminátor*, sa neschovávali v jadrových úkrytoch 29. augusta 1997. Ale tí, čo sa dopúšťajú tohto omylu, napohľad konajú ako keby videli udalosti filmu odohrať sa na *nejakej inej* planéte; nie na Zemi, ale na nejakom mieste podobnom Zemi.

Poviete: „Predstavme si, že postavíme veľmi múdru UI,“ a oni povedia: „Ale neviedlo toto k jadrovej vojne vo filme *Terminátor*?“ Pokiaľ viem, je to rovnaké uvažovanie, vrátane tónu hlasu, ako keby niekto povedal: „Ale neviedlo toto k jadrovej vojne na Alpha Centauri?“ alebo „Neviedlo toto k pádu talianskeho mestského štátu Piccolo v štrnástom storočí?“ Filmu neveríme, ale je dostupný. Neberieme ho ako proroctvo, ale ako ilustračný historický príklad. Bude sa história opakovať? Ktovie?

V nedávnej diskusii o Singularite niekto spomenul, že Vinge si zrejme nemyslí, že rozhrania medzi mozgom a počítačom môžu výrazne zvýšiť inteligenciu, a spomenul *Marooned in Realtime* a Tunça Blumenthala, ktorý bol najskúsenejší cestovateľ, ale nevyzeral až taký mocný. Rozhorčene som odpovedal: „Ale Tunç stratil väčšinu svojho hárdvéru! Bol zmrzačený!“ A potom som urobil myšlienkovú skúšku správnosti a pomyslel som si: Čo do *čerta* to trepem.

Nemali by sme o tejto téme argumentovať podľa nej samotnej, bez ohľadu na to, ako Vinge vykreslil svoje postavy? Tunç Blumenthal nie je „zmrzačený“, on je *vymyslený*. Mohol som povedať: „Vinge sa rozhodol vykresliť Tunça ako zmrzačeného z dôvodov, ktoré môžu ale nemusia súvisieť s jeho osobnými predpoveďami“ a to by dalo jeho autorskému hlasu primeranú váhu indície. Nemôže povedať: „Tunç bol zmrzačený.“ Ohľadom Tunça Blumenthala neexistuje žiadne *bol*.

Úmyselne som navrchu tohto článku ponechal chybu, ktorú som urobil pri svojom prvom náčrte: „Iní bránia svoje použitie daného *príkladu* a nesúhlasia, že je to omyl.“ Ale *Matrix nie je* príklad!

Blízky omyl je logický klam argumentovania imaginárnou indíciou: „Pozri, keby si *šiel* na koniec dúhy, *našiel* by si tam hrniec zlata – čo dokazuje, že mám pravdu!“ (Aktualizovanie na základe indície, ktorá bola predpovedaná ale nie pozorovaná je matematický zrkadlový obraz skreslenia spätného pohľadu.)

Mozog má veľa mechanizmov na zovšeobecňovanie na základe pozorovania, nielen heuristiku dostupnosti. Vidíte tri zebry, vytvoríte si kategóriu „zebra“ a táto kategória predstavuje automatický vnemový úsudok. Tvory konského tvaru s bielymi a čiernymi pruhmi sa klasifikujú ako „zebry“, preto sú rýchle a dobré na jedenie; očakávame, že budú podobné zvyšným pozorovaným zebrám.

Ľudia vidia (na obrazovke) troch Borgov, ich mozgy si automaticky vytvoria kategóriu „Borgovia“ a automaticky usudzujú, že ľudia s rozhraním medzi mozgom a počítačom budú z množiny „Borgovia“ a budú podobní pozorovaným Borgom: chladní, bezcitní, oblečení v čiernej koži, kráčajúci ťažkými mechanickými krokmi. Novinári neveria, že budúcnosť *bude* obsahovať Borgov – neveria, že *Star Trek* je proroctvo. Ale keď niekto hovorí o rozhraniach medzi mozgom a počítačom, myslia si: „Bude budúcnosť obsahovať Borgov?“ Nie: „Ako viem, že počítačom sprostredkovaná telepatia urobí ľudí menej milými?“ Nie: „Nikdy som nevidel žiadneho Borga a takisto ani nikto iný.“ Nie: „Vytváram si rasový predsudok na základe *doslova* nulovej skúsenosti.“

Ako povedal George Orwell o klišé:[107]

107 Orwell, „Politics and the English Language.“

> Predovšetkým treba nechať význam, nech si vyberie slovo, a nie naopak... Keď myslíte na niečo abstraktné, máte väčší sklon začať od slov, a pokiaľ nevynaložíte vedomé úsilie, aby ste tomu zabránili, existujúci dialekt vtrhne a vykoná svoje dielo za vás, na úkor zahmlenia alebo priam zmenenia toho, čo ste mysleli.

Odhadujem, že *najškodlivejšia* stránka používania predstavivosti iných autorov je to, že odučí ľudí používať svoju vlastnú. Ako povedal Robert Pirsig:[108]

> Bola zablokovaná, pretože sa pokúšala písomne zopakovať veci, ktoré už počula, tak ako sa on počas prvého dňa pokúšal zopakovať veci, ktoré sa už rozhodol povedať. Nedokázala vymyslieť nič, čo by sa dalo napísať o Bozemane, pretože si nedokázala spomenúť na nič počuté, čo by stálo za zopakovanie. Zvláštne si neuvedomovala, že by sa mohla pozrieť a vidieť čerstvo sama za seba, a potom písať, bez sledovania v prvom rade, čo už bolo povedané.

Zapamätané fikcie vtrhnú a urobia vaše myslenie za vás; sú náhradou *videnia* – najnebezpečnejšou zo všetkých skratiek.

*→

95. Cnosť presnosti

> Čo je pravda o jednom jablku, nemusí byť pravda o druhom jablku; preto možno o jednom jablku povedať viac než o všetkých jablkách na svete.
>
> --Dvanásť cností rozumnosti

V rámci svojej profesie ľudia rozumejú, aká dôležitá je presnosť; automechanik pozná rozdiel medzi karburátorom a radiátorom a nemyslí na ne obe ako na „súčiastky do auta“. Lovec-zberač pozná rozdiel medzi levom a panterom. Upratovač neumýva podlahu čistidlom na okná, aj keď tie fľaše pripadajú podobné tomu, kto toto umenie neovláda.

Mimo svojej profesie ľudia často robia tú chybu, že sa snažia čo najviac rozšíriť zmysel slova, aby pokrylo čo najväčšie územie. Nie je snáď slávnostnejšie, múdrejšie, pôsobivejšie rozprávať o *všetkých* jablkách na svete? O koľko vznešenejšie musí byť *vysvetliť ľudské myslenie vo všeobecnosti* než sa rozptyľovať menšími otázkami, napríklad ako ľudia vymýšľajú techniky na riešenie Rubikovej kocky. Veru, zriedkavo sa zdá potrebné všímať si *konkrétne* otázky; nie je azda všeobecná teória sama osebe dostatočne hodnotným úspechom?

Zvedaví ľudia majú vo zvyku zdvihnúť na pobreží jeden kamienok z milióna a vidieť na ňom niečo nové, niečo zaujímavé, niečo odlišné. Nazvete tieto kamienky „diamanty“ a pýtate sa, čo by na nich mohlo byť špeciálne – aké vnútorné kvality môžu mať spoločné okrem lesku, ktorý ste si všimli ako prvý. A potom príde niekto iný a povie: „Prečo nenazvať diamantom aj *tento* kamienok? A čo tento, a čo tento?“ Sú nadšení a myslia to dobre. Pretože nazývať niektoré kamienky „diamanty“ a iné nie vyzerá nedemokraticky, exkluzívne, elitársky a neholisticky. Vyzerá to... *obmedzene*... ak sa neurazíte. Nie *otvorene*, nie *všeobjímajúco*, nie *spoločensky*.

Možno si myslíte, že je poetické dať jednému slovu mnoho významov a šíriť tak odtiene konotácií navôkol. Ale aj básnik, ak je dobrým básnikom, sa musí naučiť vidieť svet presne. Nestačí prirovnať lásku ku kvetu. Horúca žiarlivá nekonzumovaná láska nie je to isté ako láska desaťročia zosobášeného páru. Ak potrebujete kvet, ktorý symbolizuje žiarlivú lásku, musíte ísť do záhrady hľadať a všímať si drobné rozdiely – nájsť kvetinu s opojnou vôňou, jasnými farbami a tŕňmi. Aj keď máte v úmysle dávať významom odtiene a šíriť konotácie, musíte pozorne sledovať, ktoré presne významy odtieňujete a konotujete.

---

108 Pirsig, *Zen and the Art of Motorcycle Maintenance.*

→ http://lesswrong.com/lw/k9/the_logical_fallacy_of_generalization_from/

Je nevyhnutnou súčasťou racionalistovho umenia – a dokonca aj básnikovho umenia! – úzko sa sústrediť na nezvyčajné kamienky, ktoré majú nejakú zvláštnu vlastnosť. A pozerať na detaily, ktoré majú tieto kamienky – ale iba tieto kamienky! – navzájom spoločné. Toto nie je hriech.

Je absolútne v poriadku, ak moderní evoluční biológovia vysvetľujú *iba* vzory živých tvorov a nie „evolúciu“ hviezd alebo „evolúciu“ technológie. Žiaľ, niektorí nešťastníci používajú slovo „evolúcia“ zároveň na označenie prirodzene vybraných vzorov replikujúceho sa života *aj* na čisto náhodnú štruktúru hviezd *aj* na inteligentne navrhnutú štruktúru technológie. A ako všetci vieme, keď ľudia používajú to isté slovo, musí to byť tá istá vec. Mali by ste automaticky zovšeobecniť všetko, čo si myslíte, že viete o biologickej evolúcii, aj na technológiu. Každý, kto hovorí opak, musí byť nezmyselne pedantný. Nie je predsa možné, aby vaše priepastné ignorantstvo ohľadom modernej evolučnej teórie bolo také veľké, že nepoznáte rozdiel medzi karburátorom a radiátorom. To je nemysliteľné. Nie, to ten *druhý* – viete, ten čo to naozaj matematicky študoval – je príliš hlúpy na to, aby videl súvislosti.

A čo už môže byť cnostnejšie než vidieť súvislosti? Nepochybne, najmúdrejšími spomedzi všetkých ľudí sú guruovia New Age, ktorí hovoria: „Všetko súvisí so všetkým.“ Ak to niekedy vyslovíte nahlas, mali by ste urobiť pauzu, aby sa každý stihol zotaviť z čistého šoku z tejto Hlbokej Múdrosti.

Existuje triviálne mapovanie medzi grafom a jeho komplementom. Úplný graf, s hranou medzi každými dvoma vrcholmi, obsahuje rovnaké množstvo informácie ako graf, ktorý nemá žiadne hrany. Dôležité grafy sú tie, v ktorých niektoré veci *nie* sú spojené s niektorými inými vecami.

Keď sa nevedomí snažia vyzerať hlboko, vykresľujú nekonečné slovné porovnania medzi touto témou a tamtou témou, čo je ako toto, čo je ako tamto; dokiaľ ich graf nie je úplne prepojený a zároveň úplne zbytočný. Liekom sú konkrétne vedomosti a štúdium do hĺbky. Keď rozumiete podrobnostiam vecí, vidíte, že *nie* sú jedna ako druhá a začnete nadšene *odoberať* hrany zo svojho grafu.

Podobne, dôležité pojmy sú tie, ktoré nezahŕňajú všetko možné vo vesmíre. Dobré hypotézy dokážu vysvetliť niektoré možné výsledky, ale nie iné.

Je absolútne v poriadku, že Isaac Newton vysvetlil *iba* gravitáciu, *iba* spôsob, ako veci padajú dole – a ako planéty obiehajú okolo Slnka a ako Mesiac spôsobuje príliv – ale *nie* úlohu peňazí v ľudskej spoločnosti, ani ako srdce pumpuje krv. Uškŕňať sa nad presnosťou je spomienkou na starých Grékov, ktorí si mysleli, že vyjsť von a naozaj sa na veci *pozrieť* je manuálna práca, a manuálna práca je pre otrokov.

Ako vyjadril Platón (v *Republike, knihe VII*):[109]

> Keby niekto zaklonil hlavu a učil by sa niečo pozeraním na rôzne vzory na strope, iste by ste si mysleli, že uvažuje svojím rozumom, zatiaľ čo on by iba pozeral svojimi očami... Verím, že žiadne štúdium neupriamuje dušu pozerať nahor, okrem toho, ktoré sa týka skutočnosti bytia a toho, čo nevidíme. Či už niekto pozerá nahor s otvorenými ústami, alebo pozerá nadol so zatvorenými ústami, ak sú to veci zmyslov, o ktorých sa snaží niečo naučiť, vyhlasujem, že sa nikdy nenaučí, pretože žiadna z týchto vecí nepripúšťa poznanie: Hovorím, že jeho duša pozerá nadol, nie nahor, aj keby ležal na chrbte na zemi či na mori!

Mnohí dnes robia podobnú chybu a myslia si, že úzke pojmy sú nehodné a nevznešené a nefilozofické, asi ako povedzme ísť von a pozerať sa na veci – konanie hodné iba spodiny. Avšak racionalisti – a aj básnici – potrebujú úzke pojmy na vyjadrenie presných myšlienok; potrebujú pojmy, ktoré zahŕňajú iba niektoré veci a vylučujú iné. Nie je nič zlé na sústredení myšlienok, zúžení pojmov, vylúčení možností a zaostrení svojich tvrdení. Naozaj, nie je! Ak budú vaše slová príliš široké, skončíte s niečím, čo nie je pravda, ba ani dobrá poézia.

A radšej mi ANI NESPOMÍNAJTE ľudí, ktorí si myslia, že Wikipédia je „umelá inteligencia“, vynález LSD bola „singularita“ alebo že korporácie sú „superinteligentné“!

* →

109 Plato, *Great Dialogues of Plato*, ed. Eric H. Warmington and Philip G. Rouse (Signet Classic, 1999).
→ http://lesswrong.com/lw/ic/the_virtue_of_narrowness/

96. Ako vyzerať (a byť) hlboký

Nedávno som sa zúčastnil diskusnej skupiny, ktorej téma pri danom stretnutí bola Smrť. Vyvolalo to hlboké emócie. Myslím si, že zo všetkých obedov v Silicon Valley, ktorých som sa zúčastnil, tento bol najúprimnejší; ľudia hovorili o smrti v rodine, o smrti priateľov, čo si mysleli o svojej vlastnej smrti. Ľudia sa navzájom naozaj počúvali. Kiež by som vedel tieto okolnosti spoľahlivo reprodukovať.

Bol som jediný prítomný transhumanista, a dával som si obrovský pozor, aby som to každému nestrkal pred nos. („Fanatik je človek, ktorý nedokáže zmeniť názor a nechce zmeniť tému.“ Ja sa snažím sa byť schopný aspoň zmeniť tému.) Nie je prekvapivé, že ľudia hovorili o význame, ktorý smrť dáva životu, alebo ako je smrť v skutočnosti maskovaným požehnaním. Ja som však, veľmi opatrne, vysvetlil, že transhumanisti sú vo všeobecnosti pozitívni voči životu, ale nemajú radi smrť.

Dodatočne za mnou prišlo pár ľudí a povedalo mi, že som veľmi „hlboký“. No áno, som, ale vďaka tomuto som začal rozmýšľať, čo spôsobuje, že ľudia *vyzerajú* hlbokí.

V jednom bode rozhovoru jedna žena povedala, že myšlienka na smrť ju vedie k tomu, aby bola k ľuďom milá, pretože nikdy nevie, či ich nevidí poslednýkrát. „Keď môžem o niekom povedať niečo pekné,“ povedala, „poviem mu to hneď a nečakám.“

„To je krásna myšlienka,“ povedal som, „a aj keby z vás jedného dňa spadla hrozba smrti, dúfam, že to budete robiť naďalej...“

Dodatočne, táto žena bola jedna z tých, čo mi povedali, že som hlboký.

V inom bode diskusie jeden muž hovoril o tom, že smrť má nejakú výhodu X, už si presne nespomínam čo. A ja som povedal: „Viete, pri ľudskej povahe, keby ľudia dostávali každý týždeň po hlave baseballovou pálkou, čoskoro by si vymysleli dôvody, prečo dostávať po hlave baseballovou pálkou je dobrá vec. Ale keby ste vzali niekoho, kto nedostával po hlave baseballovou pálkou, a opýtali by ste sa ho, či by chcel, povedal by, že nie. Myslím si, že keby ste vzali niekoho, kto bol nesmrteľný, a opýtali by ste sa ho, či by chcel zomrieť kvôli výhode X, povedal by, že nie.“

Dodatočne mi tento muž povedal, že som hlboký.

Korelácia nie je kauzalita. Možno som v ten deň jednoducho hovoril hlbším hlasom, takže to znelo múdrejšie.

Ale mám podozrenie, že som im pripadal „hlboký“, pretože som sústavne porušoval ich uložené vzory „hlbokej múdrosti“ spôsobom, ktorý hneď dával zmysel.

Existuje stereotyp Hlbokej Múdrosti. Smrť: doplňte vzor: „Smrť dáva životu zmysel.“ Každý pozná túto štandardnú Hlboko Múdru odpoveď. A preto získava určité vlastnosti signálu na potlesk. Keď ju povie, ľudia možno prikývnu, pretože ich mozog si doplní vzor a oni vedia, že majú prikývnuť. Možno dokonca povedia: „Aká hlboká múdrosť!“, azda v nádeji, že budú sami považovaní za hlbokých. Ale nebudú *prekvapení*; nepočuli nič mimo vychodených koľají; nepočuli nič, čo by si nemohli pomyslieť sami. Dalo by sa to nazvať vierou v múdrosť – myšlienka je označená ako „hlboko múdra“, a je to doplnený štandardný vzor pre „hlbokú múdrosť“, ale neprináša žiaden zážitok vhľadu.

Ľudia, ktorí sa *snažia vyzerať* Hlboko Múdri často nakoniec vyzerajú prázdni, ako ozvena, pretože sa snažia vyzerať Hlboko Múdri namiesto optimalizovania.

Koľko rozmýšľania som potreboval ja v tomto procese vyzerania hlboký? Ľudský mozog beží iba rýchlosťou 100 Hz a ja som odpovedal v reálnom čase, takže väčšina práce musela byť vypočítaná vopred. Časť, ktorú som prežíval ako náročnú, bola vyberanie odpovede pochopiteľnej v jednom inferenčnom kroku a potom jej formulovanie pre maximálny dopad.

Z filozofického hľadiska bola takmer všetka moja práca už urobená. Doplňte vzor: Existujúca situácia X je naozaj odôvodnená, pretože má výhodu Y: „Klam naturalizmu?“ / „Skreslenie status quo?“ / „Mohli by sme mať Y bez X?“ / „Keby sme nikdy predtým nepočuli o X, vybrali by sme si to dobrovoľne, aby sme dostali Y?“ Myslím si, že je férové povedať, že vykonávam tieto myšlienkové vzory asi na rovnakej úrovni automatiky ako dýcham. Napokon, väčšina ľudského myslenia musí byť vyhľadávanie uložených údajov, ak má mozog vôbec fungovať.

A tiež som už mal rozvinutú filozofiu transhumanizmu. Transhumanizmus má tiež uložené myšlienky o smrti. Smrť: doplňte vzor: „Smrť je nezmyselná tragédia, ktorú si ľudia racionalizujú.“ Toto bola neštandardná uložená informácia; taká, ktorú moji poslucháči nepoznali. Mal som niekoľko príležitostí použiť neštandardnú uloženú informáciu, a pretože to všetko boli časti rozvinutej filozofie transhumanizmu, všetky viditeľne patrili k rovnakej téme. To spôsobilo, že som vyzeral *súvisle*, nielen originálne.

Predpokladám, že toto je dôvod, prečo východná filozofia pripadá hlboká západným ľuďom – má neštandardnú ale súvislú uloženú Hlbokú Múdrosť. Zrkadlovo, v dielach japonskej fikcie človek občas nájde kresťanov vykreslených ako studnice hlbokej múdrosti a mystických tajomstiev. (A občas nie.)

Ak si správne spomínam, jeden ekonóm raz poznamenal, že laická verejnosť má tak málo skúseností so štandardnou ekonomikou, že keď ho raz pozvali vystupovať v televízii, potreboval im iba zopakovať ekonómiu pre prvákov, aby vyznel ako brilantný originálny mysliteľ.

Kritické bolo aj to, že moji poslucháči mohli *ihneď* vidieť, že moje odpovede dávajú zmysel. Mohli ale nemuseli súhlasiť s myšlienkou, ale nevyznela im ako úplne od veci. Poznám transhumanistov, ktorí nedokážu vyzerať hlbokí, pretože nedokážu zohľadniť, čo ich poslucháč ešte nevie. Ak chcete znieť hlboko, nikdy nehovorte nič, čo je vzdialené viac než jeden inferenčný krok od súčasného myšlienkového stavu vášho poslucháča. Tak to skrátka je.

Ak chcete *vyzerať* hlboký, študujte neštandardné filozofie. Vyhľadávajte diskusie na témy, ktoré vám dajú príležitosť vyzerať hlboký. Urobte svoje filozofické myslenie v predstihu, aby ste sa mohli sústrediť na dobré vysvetľovanie. Ale najmä trénujte držanie sa v hraniciach jedného inferenčného kroku.

Aby ste *boli* hlbokí, myslite sám za seba o „múdrych“ alebo dôležitých alebo emocionálne nabitých témach. Myslieť sám za seba nie je to isté ako prísť s nezvyčajnou odpoveďou. Znamená to vidieť veci sám, nie iba nechať svoj mozog doplniť vzor. Ak sa nezastavíte pri prvej odpovedi a vyhodíte odpovede, ktoré vyzerajú nejasne neuspokojivo, časom vaše myšlienky vytvoria koherentný celok, pochádzajúci z jedného zdroja, z vás, namiesto úlomkov opakovania záverov druhých ľudí.

*→

97. Svoje názory meníme zriedkavejšie, než si myslíme

> Počas posledných pár rokov sme diskrétne oslovovali kolegov, ktorí stáli pred voľbou medzi ponukami práce, a žiadali sme ich o odhad pravdepodobnosti, že si vyberú tú alebo onú prácu. Priemerná istota predpovedanej voľby bola skromných 66%, ale iba 1 z 24 respondentov si vybral možnosť, ktorej pôvodne priradil nižšiu pravdepodobnosť, z čoho vychádza celková miera presnosti 96%.
>
> --Dale Griffin a Amos Tversky[110]

Keď som prvýkrát čítal uvedené slová – 1. augusta 2003, asi o tretej popoludní – zmenili môj spôsob myslenia. Uvedomil som si, že *akonáhle viem uhádnuť, aká bude moja odpoveď* – akonáhle viem priradiť vyššiu pravdepodobnosť rozhodnutiu sa jedným smerom oproti druhému – potom som sa, s najväčšou pravdepodobnosťou, už rozhodol. Svoje názory meníme zriedkavejšie než si myslíme. Väčšinu času by sme dokázali uhádnuť svoju vlastnú odpoveď do pol sekundy po počutí otázky.

Ako rýchlo uplynie tento nepovšimnutý okamih, keď ešte nevieme uhádnuť, aká bude naša odpoveď; to malé okno príležitosti, aby inteligencia konala. V otázkach výberu, ako aj v otázkach faktu.

Princíp spodného riadku je, že iba skutočné príčiny vašich názorov určujú vašu efektivitu ako racionalistu. Akonáhle je váš názor pevne daný, žiadne množstvo argumentov nezmení jeho pravdivostnú hodnotu; akonáhle je vaše rozhodnutie pevne dané, žiadne množstvo argumentov nezmení jeho dôsledky.

→ http://lesswrong.com/lw/k8/how_to_seem_and_be_deep/

110 Dale Griffin and Amos Tversky, „The Weighing of Evidence and the Determinants of Confidence,“ [Zvažovanie indície a determinanty istoty] *Cognitive Psychology* 24, no. 3 (1992): 411–435, doi:10.1016/0010-0285(92)90013-R.

Možno si myslíte, že môžete dôjsť k názoru alebo rozhodnutiu pomocou nerozumných prostriedkov, potom ho skúsite zdôvodniť, a keď zistíte, že ho nedokážete zdôvodniť, odmietnete ho.

Lenže my svoje názory meníme zriedkavejšie – *omnoho* zriedkavejšie – než si myslíme.

Som si istý, že si dokážete spomenúť aspoň na jednu okolnosť vo svojom živote, keď ste zmenili názor. To dokáže každý. Ale čo všetky tie okolnosti vo vašom živote, keď ste svoj názor nezmenili? Sú aj tie dostupné vo vašom heuristickom odhade vlastnej kompetencie?

Uprostred skreslenia spätného pohľadu, falošnej kauzality, pozitívneho skreslenia, ukotvovania/primingu, a tak ďalej, a tak ďalej, a najmä obávaného sklonu potvrdzovať, akonáhle sa raz nejaká myšlienka dostane do vašej hlavy, pravdepodobne tam aj zostane.

*→

98. Odložte navrhovanie riešení

Z knihy Robyna Dawesa *Rozumná voľba v neistom svete*.[111] Pridané zvýraznenie.

> Norman R. F. Maier si všimol, že keď skupina čelí problému, jej členovia majú prirodzený sklon na začiatku diskusie o probléme navrhovať možné riešenia. V dôsledku toho sa interakcia skupiny zameriava na výhody a nedostatky navrhovaných riešení, ľudia emocionálne priľnú k riešeniam, ktoré sami navrhli, a lepšie riešenia už sa nenavrhujú. Maier vydal nariadenie ako zlepšiť skupinové riešenie problémov: **„Nenavrhujte riešenia, dokiaľ nebol problém nebol prediskutovaný tak dôkladne, ako sa len dá bez ich navrhovania.“** Je jednoduché ukázať, že toto nariadenie funguje v kontextoch, kde existujú objektívne definované dobré riešenia problémov.
>
> Maier vymyslel nasledujúci pokus „hrania rolí“, aby demonštroval tento bod. Pri bežiacom páse pracujú traja zamestnanci s rôznymi úrovňami zručnosti. Striedajú sa medzi troma činnosťami, ktoré vyžadujú rôznu úroveň zručnosti, pretože ten najšikovnejší – ktorý je zároveň najdominantnejší – má silnú motiváciu vyhnúť sa nude. V kontraste s ním si ten najmenej šikovný pracovník uvedomuje, že tie zložitejšie úkony nerobí rovnako dobre ako zvyšní dvaja, ale súhlasil so striedaním kvôli dominancii jeho šikovného kolegu. „Odborník na efektivitu“ si všimol, že keby najšikovnejší zamestnanec dostal tú najnáročnejšiu prácu a ten najmenej šikovný tú najmenej náročnú, produktivita by sa mohla zvýšiť o 20 % a tak odborník navrhuje, aby sa prestali striedať. Traja robotníci a štvrtá osoba, ktorá má pridelenú rolu predáka, majú za úlohu diskutovať o odporúčaniach odborníka. Niektoré skupiny dostali Maierovo nariadenie, aby nediskutovali o riešeniach, kým dôkladne neprediskutujú samotný problém, iné skupiny toto nariadenie nedostali. Tí, ktorí tento pokyn nedostali, okamžite začali argumentovať o dôležitosti produktivity verzus autonómie robotníka a o vyhýbaní sa nude. Skupiny s daným pokynom s omnoho väčšou pravdepodobnosťou objavili riešenie, že dvaja šikovnejší robotníci sa môžu striedať, zatiaľ čo ten najmenej šikovný zostane pri najmenej náročnej práci – riešenie, ktoré viedlo k zvýšeniu produktivity o 19%.
>
> Často som používal tento pokyn, keď som viedol skupiny – **najmä keď čelili veľmi ťažkému problému, čo je situácia, kde majú členovia skupiny najväčší sklon ihneď navrhovať riešenia.** Aj keď som nemal objektívne kritériá na posúdenie ako kvalitne skupina vyriešila problém, zdalo sa mi, že Maierov pokyn podporuje lepšie riešenia problémov.

To je také pravdivé, že to ani nie je smiešne. A je to horšie a horšie čím ťažší je daný problém. Vezmite si napríklad umelú inteligenciu. Prekvapujúce množstvo ľudí, s ktorými sa stretnem, vie napohľad úplne presne, ako sa má postaviť všeobecná umelá inteligencia, hoci nevedia napríklad povedať, ako postaviť optický rozoznávač znakov alebo kolaboratívny filtrujúci systém (omnoho ľahšie

→ http://lesswrong.com/lw/jx/we_change_our_minds_less_often_than_we_think/

111 Dawes, *Rational Choice in An Uncertain World*, 55–56.

problémy). Pokiaľ ide o zostrojenie umelej inteligencie s pozitívnym dopadom na svet – Priateľskej UI, voľne povedané – nuž *ten* problém je taký neuveriteľne zložitý, že *väčšina* vyrieši celú túto tému za 15 sekúnd. *Dajte mi pokoj*.

Tento problém zďaleka nie je jedinečný pre UI. Fyzici stretávajú mnoho nefyzikov, ktorí majú svoje vlastné fyzikálne teórie; ekonómovia počujú mnoho úžasných nových ekonomických teórií. Ak ste evolučný biológ, každý, koho stretnete, dokáže okamžite vyriešiť ľubovoľný otvorený problém vo vašej oblasti, zvyčajne predpokladajúc skupinovú selekciu. A tak ďalej.

Maierova rada zdôrazňuje princíp spodného riadku, že efektivitu nášho rozhodovania určuje iba to, ktoré indície a postupy sme použili, keď sme prvýkrát došli k nášmu riešeniu – keď už napíšete spodný riadok, je neskoro písať nad neho ďalšie dôvody. Ak urobíte svoje rozhodnutie príliš zavčasu, naozaj bude založené na veľmi malom množstve rozmýšľania, bez ohľadu na to, koľko úžasných argumentov vymyslíte dodatočne.

Ďalej uvážte, že svoje názory meníme zriedkavejšie než si myslíme: 24 ľudí priradilo v priemere pravdepodobnosť 66 %, že si v budúcnosti vyberú možnosť, ktorá im vtedy pripadala pravdepodobnejšia, ale iba 1 z 24 si naozaj vybral možnosť, ktorá mu vtedy pripadala menej pravdepodobná. **Akonáhle viete uhádnuť, aká bude vaša odpoveď, už ste sa pravdepodobne rozhodli.** Ak dokážete uhádnuť svoju odpoveď pol sekundy po tom, čo počujete otázku, mali ste iba pol sekundy na to, aby ste boli inteligentní. To nie je veľa času.

Tradičná rozumnosť zdôrazňuje *falzifikáciu* – schopnosť *vzdať sa* pôvodného názoru, keď sa stretnete s jasnou indíciou proti nemu. Ale akonáhle sa myšlienka dostane do vašej hlavy, pravdepodobne bude vyžadovať príliš veľa indícií opäť ju odtiaľ dostať. Čo je horšie, niekedy nemáme luxus prevahy indícií.

Domnievam sa, že omnoho mocnejšia (a náročnejšia) metóda je *odložiť myslenie na odpoveď*. Zastaviť sa, predĺžiť tú chvíľku, keď ešte nevieme uhádnuť, aká bude naša odpoveď; čím svojej inteligencii doprajeme viac času.

Aj pol minúty by bolo zlepšenie oproti polsekunde.

*→

99. Klam pôvodu

V zoznamoch logických chýb nájdete „klam pôvodu“ - chybné napádanie názoru na základe príčin, prečo tomu niekto verí.

To je na prvý pohľad veľmi zvláštna myšlienka – ak príčiny názoru neovplyvňujú jeho systematickú spoľahlivosť, tak čo teda? Ak nám Deep Blue odporučí ťah v šachu, budeme mu dôverovať na základe nášho porozumenia *kódu*, ktorý prehľadáva herný strom, keďže tento herný strom nedokážeme vyhodnotiť sami. Čo by označovalo nejaký pravdepodobnostný argument ako „rozumný“, ak nie to, že bol vytvorený nejakým systematicky spoľahlivým procesom?

Články o klame pôvodu vám povedia, že uvažovanie o pôvode nie je vždy chyba – že pôvod indície *môže* byť relevantný pre jej vyhodnotenie, napríklad v prípade dôveryhodného odborníka. Ale inokedy, hovoria tieto články, to *je* omyl; chemik Kekulé prvýkrát uvidel prstencovú štruktúru benzénu vo sne, ale to neznamená, že tomto názoru nesmieme nikdy dôverovať.

Takže niekedy je klam pôvodu omylom a niekedy nie?

Klam pôvodu je formálne omylom, pretože *pôvodná príčina* názoru nie je to isté ako jeho *súčasný stav zdôvodniteľnosti*, súhrn všetkých *dnes* známych argumentov za a proti.

Svoje názory však meníme zriedkavejšie než si myslíme. Obvinenia z pôvodu majú medzi ľudmi silu, ktorú by nemali medzi ideálnymi Bayesovcami.

Vyčistiť si myseľ je *silná heuristika*, ak nadobudnete nové podozrenie, že mnohé z vašich názorov môžu pochádzať z chybného zdroja.

→ http://lesswrong.com/lw/ka/hold_off_on_proposing_solutions/

Akonáhle sa nejaká myšlienka dostane do našich hláv, nie je vždy ľahké vykoreniť ju pomocou indícií. Vezmite si všetkých tých ľudí, ktorí vyrástli veriac Biblii; neskôr odmietli (na vedomej úrovni) predstavu, že Bibliu napísala ruka Boha; a ktorí si predsa myslia, že Biblia obsahuje nevyhnutnú etickú múdrosť. Zabudli si vyčistiť myseľ; robili by omnoho lepšie, keby pochybovali o všetkom, čo Biblia hovorí, *pretože to Biblia hovorí.*

Zároveň by mali v hlavne pevne držať princíp, že obrátená hlúposť nie je inteligencia; cieľom je skutočne si uvoľniť hlavu a myslieť nezávisle, nie negovať Bibliu a mať toto ako svoj algoritmus.

Akonáhle sa nejaká myšlienka dostane do vašej hlavy, budete mať sklon vidieť pre ňu podporu kamkoľvek sa pozriete – a preto, keď je pôvodný zdroj náhle v podozrení, bolo by veru veľmi múdre pochybovať o listoch, ktoré pôvodne vyrástli na tejto vetve...

Ak to dokážete! Nie je ľahké vyčistiť si myseľ. Vyžaduje to kŕčovité úsilie *naozaj znovu uvažovať*, namiesto nechania svojej mysle, aby zapadla do vzorca opakovania uložených argumentov. „Nie je to skutočná kríza viery, dokiaľ by veci nemohli rovnako ľahko ísť ľubovoľným smerom," povedal Thor Shenkel.

Mali by ste byť *mimoriadne podozrievaví*, ak máte mnoho myšlienok, ktoré vám odporučil zdroj, o ktorom teraz viete, že je nedôveryhodný, ale zhodou okolnosti sa zdá, že všetky tieto myšlienky boli správne – Biblia je tu samozrejmým archetypálnym príkladom.

Na druhej strane... existuje niečo také ako indícia dostatočne jasná, že už významne nezáleží na tom, odkiaľ táto myšlienka pôvodne prišla. Zhromažďovanie tohto druhu jasných indícií, o tom je celá veda. Nezáleží už na tom, že Kekulé prvýkrát uvidel prstencovú štruktúru benzénu v sne – nezáležalo by na tom, keby sme túto hypotézu na testovanie našli pomocou generovania náhodných obrázkov na počítači, od spiritualistu, ktorý bol odhalený ako podvodník, alebo dokonca z Biblie. Prstencová štruktúra benzénu je podporená toľkými pokusnými indíciami, že zdroj tohto námetu je nepodstatný.

V neprítomnosti takejto jasnej indície musíte venovať pozornosť pôvodným zdrojom myšlienok – dôverovať odborníkom viac než laikom, ak si ich oblasť získala rešpekt – podozrievať myšlienky, ktoré ste pôvodne získali z podozrivých zdrojov – nedôverovať tým, ktorých motívy sú nedôveryhodné, *ak* nedokážu predložiť argumenty nezávislé na ich vlastnej autorite.

Klam pôvodu je *omylom* vtedy, keď existujú zdôvodnenia *okrem* pôvodného tvrdenia faktu, ale keď sa obvinenie z pôvodu prekladá, akoby to uzavrelo tému. Hal Finney odporúča, aby sme správne odvolávanie sa na pôvod tvrdenia nazývali „heuristika pôvodu".

Niektoré dobré heuristiky (pre ľudí):

- Dávajte si pozor na všeobecné obvinenia voči názorom, ktoré sa vám nepáčia, najmä ak váš oponent tvrdí, že má aj iné dôvody okrem jednoduchej autority hovoriaceho. „Lietanie je náboženská predstava, preto bratia Wrightovci museli byť klamári" je jeden z klasicky uvádzaných príkladov.
- Rovnako si nemyslite, že dostanete dobrú informáciu o technickej téme tým, že budete múdro psychoanalyzovať zúčastnené osobnosti a ich pomýlené motívy. Ak existujú technické argumenty, majú prednosť.
- Keď vznikne nové podozrenie ohľadom niektorého z vašich základných zdrojov, naozaj by ste *mali* pochybovať o všetkých vetvách a listoch, ktoré z tohto koreňa vyrástli. Neoprávňuje vás ho priamo tieto závery vyhodiť, lebo obrátená hlúposť nie je inteligencia, ale...
- Buďte extrémne podozrievaví, ak zistíte, že stále veríte dávnym odporučeniam zdroja, ktorý ste neskôr odmietli.

* →

→ http://lesswrong.com/lw/s3/the_genetic_fallacy/

J: Špirály smrti

100. Afektívna heuristika

Afektívna heuristika je keď subjektívne vnímanie niečoho ako dobrého alebo zlého funguje ako heuristika – zdroj rýchlych vnemových úsudkov. Príjemné a nepríjemné pocity sú v strede ľudského uvažovania a afektívna heuristika prichádza s krásnymi skresleniami – niektorými z mojich najobľúbenejších.

Začnime jedným z tých pomerne menej bláznivých skreslení. Idete sa sťahovať do nového mesta a musíte prepraviť starožitné kyvadlové hodiny. V prvom prípade sú tieto starožitné hodiny darom od vašich starých rodičov k vašim 5. narodeninám. V druhom prípade sú hodiny darom od vzdialeného príbuzného a nemáte k nim žiadne zvláštne pocity. Koľko by ste boli ochotní zaplatiť za poistku, ktorá vám vyplatí 100 dolárov ak sa tieto hodiny počas prevozu stratia? Podľa Hsee a Kunreuthera pokusné osoby uvádzali ochotu zaplatiť viac než dvojnásobnú sumu v prvom prípade.[112] To môže znieť rozumne – prečo neplatiť viac za ochranu cennejšieho majetku? - kým si neuvedomíte, že poistka tie hodiny *nechráni*, iba vyplatí peniaze, ak sa hodiny stratia, a vyplatí rovnakú sumu v oboch prípadoch. (A áno, bolo uvedené, že poistka je uzavretá mimo dopravnej firmy, takže nedáva sťahovákom žiadnu zvláštnu motiváciu.)

Dobre, to však *neznie* príliš šialene. Možno by ste sa z toho dostali tvrdením, že pokusné osoby poisťovali svoje citové výsledky, nie finančné výsledky – kupovali si útechu.

A čo toto? Yamagishi ukázal, že pokusné osoby hodnotili chorobu ako nebezpečnejšiu, keď bola opísaná ako zabíjajúca 1 286 ľudí z každých 10 000 nakazených, v porovnaní s chorobou, ktorá mala úmrtnosť 24,14 %.[113] Zdá sa, že predstava tisíc mŕtvol je omnoho znepokojujúcejšia než jedna osoba, ktorá má väčšiu šancu prežiť než umrieť.

Však počkajte, bude to horšie.

Predstavte si, že sa letisko musí rozhodnúť, či použije peniaze na nákup nového zariadenia, kým kritici argumentujú, že by sa peniaze mali radšej použiť na iné stránky bezpečnosti letiska. Slovic a kol. predložili dvom skupinám pokusných osôb argumenty za a proti nákupu zariadenia, pričom reakciu merali na škále od 0 (vôbec by nepodporil) po 20 (veľmi silná podpora).[114] Jedna skupina mala zariadenie opísané tak, že zachráni 150 životov. Druhá skupina mala zariadenie opísané tak, že zachráni 98 % zo 150 životov. Hypotéza za týmto experimentom znela, že 150 životov znie neurčito dobre – je to veľa? málo? - zatiaľ čo zachrániť 98 % niečoho je jasne veľmi dobré, pretože 98 % je veľmi blízko hornej hranici percentuálnej škály. A hľa, záchrana 150 životov mala priemernú podporu 10,4, kým záchrana 98 % zo 150 životov mala priemernú podporu 13,6.

Alebo si vezmite správu od Denes-Rajovej a Epsteina:[115] Pokusné osoby, ktoré dostali možnosť vyhrať 1 dolár vždy, keď si náhodne vytiahnu červený želé cukrík z misky, často dávali prednosť miske, kde bolo červených cukríkov viac ale v menšom pomere. Napríklad 7 zo 100 malo prednosť pred 1 z 10.

Podľa Denes-Rajovej a Epsteina tieto pokusné osoby neskôr uvádzali, že hoci vedeli, že pravdepodobnosti sú proti nim, cítili, že majú väčšiu šancu, keď je viac červených cukríkov. To môže pripadať bláznivé vám, ó Štatisticky Sofistikovaný Čitateľ, ale keď sa nad tým pozornejšie zamyslíte, uvedomíte si, že to dokonale dáva zmysel. Pravdepodobnosť 7 % oproti pravdepodobnosti 10 % môže

112 Christopher K. Hsee and Howard C. Kunreuther, „The Affection Effect in Insurance Decisions," [Účinok afektu pri poistných rozhodnutiach] *Journal of Risk and Uncertainty* 20 (2 2000): 141–159, doi:10.1023/A:1007876907268.

113 Kimihiko Yamagishi, „When a 12.86% Mortality Is More Dangerous than 24.14%: Implications for Risk Communication," [Kedy je úmrtnosť 12,86 % nebezpečnejšia než 24,14 %: Dôsledky pre komunikáciu rizika.] *Applied Cognitive Psychology* 11 (6 1997): 461–554.

114 Paul Slovic et al., „Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics," *Journal of Socio-Economics* 31, no. 4 (2002): 329–342, doi:10.1016/S1053-5357(02)00174-9.

115 Veronika Denes-Raj and Seymour Epstein, „Conflict between Intuitive and Rational Processing: When People Behave against Their Better Judgment," [Konflikt medzi intuitívnym a rozumovým spracovaním: Keď sa ľudia správajú v rozpore so svojím úsudkom.] *Journal of Personality and Social Psychology* 66 (5 1994): 819–829, doi:10.1037/0022-3514.66.5.819.

byť zlá správa, ale je viac než vyvážená väčším počtom červených cukríkov. Je to horšia pravdepodobnosť, to áno, ale aj tak máte väčšiu šancu *vyhrať*, jasné? Mali by ste nad touto myšlienkou meditovať dokiaľ nedosiahnete osvietenie ohľadom toho, ako väčšina tejto planéty uvažuje o pravdepodobnosti.

Finucane a kol. zistil, že pre jadrové reaktory, zemný plyn a konzervačné látky v potravinách predkladať ľuďom informáciu o vysokých výnosoch spôsobilo, že vnímali riziká ako nižšie; predkladať ľuďom informáciu o vyššom riziku spôsobilo, že vnímali výnosy ako nižšie; a tak ďalej vo všetkých kvadrantoch.[116] Ľudia zmiešavajú svoje úsudky o konkrétnych dobrých alebo zlých stránkach niečoho do celkového dobrého alebo zlého pocitu z danej veci.

Finucane a kol. ďalej zistili, že časový tlak výrazne *zvyšoval* nepriamo úmerný vzťah medzi vnímaným rizikom a vnímaným výnosom, čo je v súlade so všeobecným zistením, že časový tlak, slabé informácie alebo rozptyľovanie všetky zvyšujú dominanciu vnemových heuristík nad analytickou rozvahou.

Ganzach našiel ten istý efekt v oblasti financií.[117] Podľa bežnej ekonomickej teórie by výnos a riziko mali korelovať *pozitívne* – inými slovami, ľudia si priplácajú za bezpečnejšie investície, čo znižuje ich výnos; akcie dávajú väčšie výnosy než dlhopisy, ale majú primerane vyššie riziko. Pri hodnotení akcií *známych* firiem mali úsudky analytikov ohľadom rizika a výnosov priamu koreláciu, ako sa bežne predpovedalo. Pri hodnotení akcií *neznámych* firiem však analytici mali sklon hodnotiť akcie akoby boli všeobecne dobré alebo všeobecne zlé – nízke riziko a vysoký výnos, alebo vysoké riziko a nízky výnos.

Na ďalšie čítanie odporúčam skvelý súhrnný Slovicov článok: „Rozumní konatelia alebo rozumní hlupáci: Dôsledky afektívnej heuristiky pre behaviorálnu ekonómiu.“[118]

101. Vyhodnotiteľnosť (a lacné vianočné nákupy)

Ako sa blíži *drahá* časť Helovínskomikulášovskovianočného obdobia, väčšinu čitateľov musí trápiť otázka:

> Drahý *Overcoming Bias*, existujú skreslenia, ktoré môžem využiť, aby som *vyzeral* štedro a nemusel *naozaj* minúť veľa peňazí?

Rád vám oznamujem, že odpoveď je áno! Podľa Hsee – v článku nazvanom „Menej je lepšie: Keď sú lacné možnosti hodnotené vyššie než drahšie možnosti“ - ak niekomu kúpite šatku za 45 dolárov, budete pravdepodobne vnímaný ako štedrejší než keď mu kúpite kabát za 55 dolárov.[119]

Toto je špeciálny prípad všeobecnejšieho javu. V predchádzajúcom pokuse, sa Hsee sa pokusných osôb pýtal, koľko by boli ochotné zaplatiť za secondhandový hudobný slovník:[120]

- Slovník A, z roku 1993, obsahuje 10 000 hesiel, vyzerá ako nový.
- Slovník B, z roku 1993, obsahuje 20 000 hesiel, má roztrhnutú obálku, ale inak vyzerá ako nový.

116 Finucane et al., „The Affect Heuristic in Judgments of Risks and Benefits.“ [Afektívna heuristika pri posudzovaní rizík a výnosov.]

117 Yoav Ganzach, „Judging Risk and Return of Financial Assets,“ [Hodnotenie rizika a výnosov finančných aktív] *Organizational Behavior and Human Decision Processes* 83, no. 2 (2000): 353–370, doi:10.1006/obhd.2000.2914.

118 Slovic et al., „Rational Actors or Rational Fools.“ [Rozumní konatelia alebo rozumní hlupáci: Dôsledky afektívnej heuristiky pre behaviorálnu ekonómiu]

→ http://lesswrong.com/lw/lg/the_affect_heuristic/

119 Christopher K. Hsee, „Less Is Better: When Low-Value Options Are Valued More Highly than High-Value Options,“ [Menej je lepšie: Keď sú lacné možnosti hodnotené vyššie než drahšie možnosti] *Behavioral Decision Making* 11 (2 1998): 107–121.

120 Christopher K. Hsee, „The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives,“ *Organizational Behavior and Human Decision Processes* 67 (3 1996): 247–257, doi:10.1006/obhd.1996.0077.

Trik bol v tom, že niektoré pokusné osoby videli oba slovníky vedľa seba, kým iné pokusné osoby videli iba *jeden* slovník...

Osoby, ktoré videli iba *jednu* z týchto možností boli ochotné zaplatiť v priemere 24 dolárov za Slovník A a v priemere 20 dolárov za Slovník B. Osoby, ktoré videli *obe* možnosti vedľa seba, boli ochotné zaplatiť 27 dolárov za Slovník B a 19 dolárov za Slovník A.

Samozrejme, počet hesiel v slovníku je dôležitejší než či má roztrhnutú obálku, prinajmenšom ak ho niekedy plánujete na niečo použiť. Ale ak vám ukážu iba jeden slovník a ten má 20 000 hesiel, číslo 20 000 veľa nepovie. Je to málo? Veľa? Kto vie? To sa *nedá vyhodnotiť*. Na druhej strane, roztrhnutá obálka – tá bije do očí. To má jasnú afektívnu hodnotu: konkrétne, zlú.

Keď ich vidíme vedľa seba, počet hesiel sa zmení z *nevyhodnotiteľného* na *vyhodnotiteľný*, pretože môžeme porovnať dva porovnateľné údaje. A akonáhle sa počet hesiel stane porovnateľným, tento faktor prebije dôležitosť roztrhnutej obálky.

Podľa Slovic a kol.: Chceli by ste radšej:[121]

1. Šancu 29/36 vyhrať 2 doláre
2. Šancu 7/36 vyhrať 9 dolárov

Zatiaľ čo priemerné *ceny* (ekvivalentné hodnoty) odhadované pri týchto možnostiach boli 1,25 a 2,11 dolára, ich priemerné hodnotenie príťažlivosti bolo 13,2 a 7,5. Ceny aj hodnotenia príťažlivosti boli získané tak, že sa pokusným osobám povedalo, že sa náhodne vyberú dve z hodnotených stávok a oni dostanú tú s vyššou cenou alebo s vyšším hodnotením príťažlivosti. (Pokusné osoby mali motív hodnotiť stávky ako príťažlivejšie alebo cenu ako vyššiu než by v skutočnosti boli ochotné zaplatiť.)

Stávka hodná viac peňazí vyzerala menej príťažlivo, klasické prevrátenie preferencií. Výskumníci si mysleli, že dolárové výhry sú skôr v súlade s odhadovanou cenou, zatiaľ čo pravdepodobnosť výhry je skôr v súlade s príťažlivosťou. Takže (povedali si výskumníci), prečo neskúsiť urobiť výhru zo stávky emocionálne ešte vypuklejšiu – afektívne hodnotiteľnejšiu – príťažlivejšiu?

Ako to urobili? Pridaním malej straty do stávky. Pôvodná stávka mala šancu 7/36 vyhrať 9 dolárov. Nová stávka mala šancu 7/36 vyhrať 9 dolárov a 29/36 prehrať 5 centov. Pri starej stávke ste implicitne hodnotili príťažlivosť 9 dolárov. Nová stávka vás nabáda hodnotiť príťažlivosť vyhrania 9 dolárov v *porovnaní s* prehraním 5 centov.

„Výsledky," hovorí Slovic a kol., „prekonali naše očakávania." V novom experimente mala jednoduchá stávka so šancou 7/36 vyhrať 9 dolárov priemerné hodnotenie príťažlivosti 9,4, zatiaľ čo zložitejšia stávka, ktorá mala navyše šancu 29/36 prehrať 5 centov mala priemerné hodnotenie príťažlivosti 14,9.

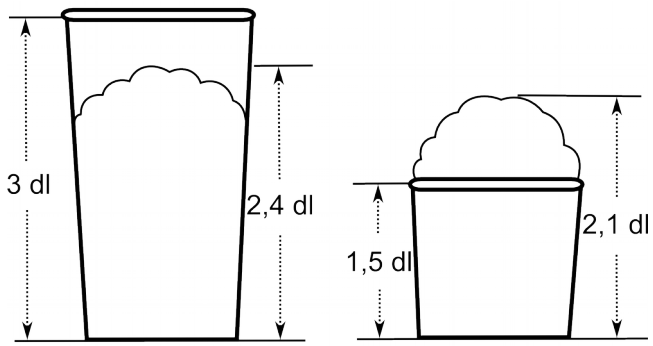
Nasledujúci pokus testoval, či pokusné osoby dajú prednosť starej stávke pred istým ziskom 2 dolárov. Iba 33 % študentov dalo prednosť starej stávke. V druhej skupine, ktorá si mala vybrať medzi istými 2 dolármi a novou stávkou (s pridanou možnosťou prehry 5 centov), až 60,8 % dalo prednosť stávke. Napokon, 9 dolárov nie je príliš príťažlivá suma peňazí, ale 9 dolárov oproti 5 centom je *mimoriadne* príťažlivý pomer výhry k prehre.

Môžete urobiť stávku atraktívnejšou tým, že k nej čisto pridáte stratu! Nie je psychológia zábavná? To je dôvod prečo nikto, to naozaj oceňuje zázračnú zložitosť ľudskej inteligencie, nechce dizajnovať UI, ktorá by bola ako človek.

Samozrejme, toto funguje iba ak pokusné osoby nevidia tie dve stávky vedľa seba.

Podobne, ktorej z nasledujúcich zmrzlín podľa vás pokusné osoby v štúdii Hsee 1998 dávali prednosť?

121 Slovic et al., „Rational Actors or Rational Fools."



Predajca V Predajca N

Prirodzene, odpoveď závisela na tom, či pokusné osoby videli iba jednu zmrzlinu alebo obe vedľa seba. Pokusné osoby, ktoré videli jednu zmrzlinu, boli ochotné zaplatiť 1,66 dolára Predajcovi V a 2,26 dolára Predajcovi N. Pokusné osoby, ktoré videli obe zmrzliny, boli ochotné zaplatiť 1,85 dolára Predajcovi V a 1,56 dolára Predajcovi N.

Čo to znamená pre vaše vianočné nákupy? Že ak miniete 400 dolárov na 16GB iPod Touch, obdarovaný uvidí najdrahší MP3 prehrávač. Ak miniete 400 dolárov na Nintendo Wii, obdarovaný uvidí najlacnejšie herné zariadenie. Čo je lepšia hodnota za vaše peniaze? Ach, ale táto otázka dáva zmysel, iba keď vidíte obe vedľa seba. *Vy* na ne budete myslieť vedľa seba, kým budete nakupovať, ale obdarovaný uvidí iba to, čo dostane.

Ak máte pevne určené množstvo peňazí na minutie – a vaším cieľom je ukázať priateľstvo, nie naozaj *pomôcť* obdarovanému – lepšie je úmyselne nekupovať hodnotu. Rozhodnite sa, koľko miniete na urobenie dojmu na obdarovaného a potom nájdite tej najmenej hodnotný predmet, ktorý toľko stojí. Čím lacnejšia je daná *kategória* predmetov, tým drahšie bude *ten konkrétny* predmet vyzerať, ak máte pevne určenú sumu na minutie. Čo si viac zapamätáte, tričko za 25 dolárov alebo sviečku za 25 dolárov?

To dáva celkom nový zmysel japonskému zvyku kupovať 50-dolárové melóny, však? Pozriete sa na to a potrasiete hlavou so slovami: „Čo to *je* s tými Japoncami?“ A predsa sú vďaka tomu vnímaní ako neuveriteľne štedrí, priam rozhadzovační, hoci minú iba 50 dolárov. Môžete minúť 200 dolárov na luxusnú večeru a nebudete vyzerať tak bohato ako by ste mohli minutím 50 dolárov za melón. Keby tak bol zvyk kupovať špáratká za 25 dolárov alebo zrnká prachu za 10 dolárov, to by sa nám mohlo prepiecť minúť ešte menej.

PS: Ak tento trik naozaj použijete, chcem vedieť, čo ste kúpili.

* →

102. Neobmedzené škály, obrovské súdne odmeny a futurizmus

„Psychofyzika“ napriek svojmu názvu je úctyhodná oblasť, ktorá spája fyzické účinky so zmyslovými účinkami. Ak vypustíte akustickú energiu do vzduchu – vydáte zvuk – *ako hlasno* to bude znieť človeku, ako funkcia akustickej energie? O koľko viac akustickej energie musíte napumpovať do vzduchu, aby to človeku znelo dvakrát hlasnejšie? Nie je to dvakrát viac; skôr asi osemkrát viac.

Akustická energia a fotóny sa merajú priamočiaro. Keď chcete zistiť, ako hlasno nejaký akustický podnet *znie*, ako jasne nejaký svetelný zdroj *vyzerá*, zvyčajne sa opýtate počúvajúceho alebo pozerajúceho. To sa dá pomocou obmedzenej škály od „veľmi tichý“ po „veľmi hlasný“ alebo „veľmi slabé“ až „veľmi jasné“. Môžete použiť aj neobmedzenú škálu, kde nula je „vôbec nepočuť“ alebo „vôbec nevidno“, ale odtiaľ možno zvyšovať bez ohraničenia. Keď používate neobmedzenú škálu, pozorovateľ zvyčajne dostane konštantný podnet, *modul*, ktorý dostane pevne dané hodnotenie. Napríklad zvuk, ktorému sa priradí číslo 10. Pozorovateľ potom môže označiť zvuk dvakrát hlasnejší ako modul napísaním 20.

→ http://lesswrong.com/lw/lh/evaluability_and_cheap_holiday_shopping/

A toto sa ukazuje ako celkom spoľahlivá technika. Ale čo sa stane, ak dáte pokusným osobám neobmedzenú škálu, ale žiaden modul? Od 0 do nekonečna a žiaden pomocný údaj s pevne danou hodnotou? Potom si vymyslia svoj vlastný modul, samozrejme. *Pomery* medzi podnetmi budú medzi pokusnými osobami naďalej spoľahlivo korelovať. Osoba A povie, že zvuk X má hlasitosť 10 a zvuk Y má hlasitosť 15. Ak osoba B povie, že zvuk X má hlasitosť 100, potom je dobrý odhad, že osoba B priradí zvuku Y hlasitosť okolo 150. Ak však neviete, čo používa osoba C ako svoj modul – svoju mierku – potom sa nedá odhadovať, čo osoba C povie na zvuk X. Možno to bude 1. Možno to bude 1000.

Ak pokusné osoby hodnotia *jediný* zvuk, na *neobmedzenej* škále, *bez* pevne daného štandardu na porovnanie, takmer *všetok* rozptyl je spôsobený svojvoľným výberom modulu, nie samotným zvukom.

„Hm," pomyslíte si, „to znie dosť ako keď porota určuje výšku odškodného. Nečudo, že je tam taký rozptyl!" Zaujímavá analógia, ale ako by ste to dokázali pokusom?

Kahneman a kol., 1998 a 1999, predložili 867 pokusným osobám spôsobilým sedieť v porote opisy súdnych prípadov (napr. dieťa, ktorému sa zapálili šaty) a požiadal ich o jedno z nasledujúceho:

1. Ohodnotiť nehoráznosť činov obvineného, na ohraničenej škále
2. Ohodnotiť stupeň nakoľko by mal obvinený byť potrestaný, na ohraničenej škále, alebo
3. Určiť výšku odškodného v dolároch.[122]

A hľa, zatiaľ čo pokusné osoby navzájom veľmi dobre korelovali v hodnoteniach nehoráznosti a hodnoteniach trestu, ich odškodné bolo od buka do buka. Napriek tomu *poradové hodnotenia* odškodného od danej osoby – poradové číslo od najnižšieho odškodného po najvyššie odškodné – dobre korelovali medzi osobami.

Ak chcete vedieť, akú časť rozptylu na škále „trestu" možno vysvetliť konkrétnym scenárom – konkrétnym súdnym prípadom, ktorý bol predložený viacerým pokusným osobám – potom odpoveď, ešte aj pre hrubé skóre, bola 0.49. Pre *poradia* dolárových odpovedí bolo predpovedané množstvo rozptylu 0,51. Pre *hrubé dolárové* údaje bol vysvetlený rozptyl iba 0,06!

Čo znamená: keby ste poznali predložený scenár – už spomínané dieťa, ktorému sa zapálili šaty – mohli by ste dobre odhadnúť hodnotenie trestu, dobre odhadnúť *poradie* odškodného v dolároch oproti iným prípadom, avšak samotná výška odškodného v dolároch by bola celkom nepredvídateľná.

Vziať medián dvanástich náhodne vybraných odpovedí tiež veľmi nepomohlo.

Určenie odškodného porotou teda nie je natoľko ekonomickým ocenením ako vyjadrením postoja – psychologickou mierou rozhorčenia vyjadrenou na neobmedzenej škále bez štandardného modulu.

Všímam si, že mnohé *futuristické predpovede* je podobne najlepšie vnímať ako vyjadrenia postoja. Vezmite si otázku: „Ako dlho potrvá, kým budeme mať UI na úrovni človeka?" Odpovede, ktoré som na toto videl, siahajú od buka do buka. Pri jednej pamätihodnej príležitosti mi človek zo štandardnej oblasti UI povedal: „Päťsto rokov." (!!)

Samotný dôvod, prečo sa čas do UI *nedá dobre predpovedať*, je téma sa samostatnú dlhú diskusiu. Ale neznamená to, že ten človek, ktorý povedal „Päťsto rokov", videl túto odpoveď v budúcnosti. Ani to číslo nemohol dostať štandardnou pochybnou metódou pomocou Moorovho zákona. Čo teda toto číslo 500 *znamenalo*?

Podľa môjho odhadu to je, ako keby ste sa pýtali: „Na škále, kde nula znamená ‚vôbec nie zložité', aký zložitý sa vám *zdá* problém UI?" Keby to bola obmedzená škála, každý príčetný respondent by označil odpoveď „extrémne ťažký" na pravom konci. Všetko vám *pripadá* extrémne ťažké, keď neviete, ako sa to robí. Ale namiesto toho je tu neobmedzená škála bez štandardného modulu. Ľudia si teda vymyslia číslo, ktoré predstavuje „extrémne zložité", čo môže byť 50, 100 alebo dokonca 500. Potom pridajú na koniec slovo „rokov" a to je ich futuristická predpoveď.

122 Daniel Kahneman, David A. Schkade, and Cass R. Sunstein, „Shared Outrage and Erratic Awards: The Psychology of Punitive Damages," [Spoločné rozhorčenie a náhodné odškodné: Psychológia odškodného] *Journal of Risk and Uncertainty* 16 (1 1998): 48–86, doi:10.1023/A:1007710408413; Daniel Kahneman, Ilana Ritov, and David Schkade, „Economic Preferences or Attitude Expressions?: An Analysis of Dollar Responses to Public Issues," [Ekonomické preferencie alebo vyjadrenia postoja? Analýza dolárových reakcií na verejné témy] *Journal of Risk and Uncertainty* 19, nos. 1–3 (1999): 203–235, doi:10.1023/A:1007835629236.

„Aký ťažký sa vám zdá problém UI?“ nie je jediná náhradná otázka. Iní odpovedajú, ako keby som sa pýtal: „Ako pozitívne sa cítite ohľadom UI?“, kde nižšie čísla znamenajú pozitívnejšie pocity a tiež pridajú na koniec slovo „rokov“. Ale ak tieto „časové odhady“ predstavujú niečo iné než vyjadrenia postojov na neobmedzenej škále bez modulu, nedokázal som určiť, čo.

⁂→

103. Efekt svätožiary

Afektívna heuristika je, keď celkový pocit dobrého alebo zlého prispieva k mnohým ďalším úsudkom, či už je to logické alebo nie, či si to uvedomujete alebo nie. Pokusné osoby, ktorým sa povedalo o výnosoch jadrovej elektrárne majú sklon hodnotiť ju ako menej rizikovú; burzoví analytici hodnotiaci akcie neznámych firiem ich hodnotia ako všeobecne dobré alebo všeobecne zlé – nízke riziko a vysoké výnosy, alebo vysoké riziko a nízke výnosy – v rozpore s bežnou ekonomickou teóriu, ktorá hovorí, že riziko a výnosy by mali pozitívne korelovať.

Efekt svätožiary je prejav afektívnej heuristiky v sociálnej psychológii. Robert Cialdini v knihe *Vplyv: Veda a prax*[123] zhŕňa:

> Výskum ukázal, že dobre vyzerajúcim jednotlivcom automaticky pripisujeme také priaznivé črty ako talent, láskavosť, čestnosť a inteligencia (prehľadová štúdia týchto indícií viď Eagly, Ashmore, Makhijani, & Longo, 1991).[124] Navyše, tieto úsudky robíme bez vedomia, že fyzická príťažlivosť hrá v tomto procese rolu. Niektoré dôsledky tohto nevedomého predpokladu, že „dobre vyzerajúci rovná sa dobrý“ ma desia. Napríklad štúdia kanadských federálnych volieb v roku 1974 zistila, že atraktívni kandidáti získali vyše dva-a-pol-krát viac hlasov ako neatraktívni kandidáti (Efran & Patterson, 1976).[125] Napriek takýmto indíciám uprednostňovania pekných politikov, nasledujúci výskum ukázal, že voliči si neuvedomujú svoje skreslenie. V skutočnosti 73 percent oslovených kanadských voličov odmietlo najsilnejším možným spôsobom, že by ich hlasy boli ovplyvnené fyzickým výzorom; iba 14 percent vôbec pripustilo možnosť takéhoto vplyvu (Efran & Patterson, 1976).[126] Voliči môžu popierať vplyv príťažlivosti na zvoliteľnosť, koľko len chcú, ale indície naďalej potvrdzujú jej problematickú prítomnosť (Budesheim & DePaola, 1994).[127]
>
> Podobný účinok sa našiel pri prijímacích pohovoroch. Podľa jednej štúdie mala dobrá upravenosť uchádzačiek v simulovanom pracovnom pohovore väčší vplyv na rozhodnutie zamestnať než pracovná kvalifikácia – napriek tomu, že osoby robiace rozhovor tvrdili, že výzor hral v ich rozhodovaní malú rolu (Mack & Rainey, 1990).[128] Výhoda príťažlivých uchádzačov sa týkala nielen zamestnania, ale aj výplaty. Ekonómovia skúmajúci americké a kanadské vzorky zistili, že príťažliví zamestnanci dostávajú v priemere o 12 až 14 percent vyššie výplaty než ich nepríťažliví spolupracovníci (Hammermesh & Biddle, 1994).[129]

→ http://lesswrong.com/lw/li/unbounded_scales_huge_jury_awards_futurism/

123 Robert B. Cialdini, *Influence: Science and Practice* (Boston: Allyn & Bacon, 2001).

124 Alice H. Eagly et al., „What Is Beautiful Is Good, But . . . A Meta-analytic Review of Research on the Physical Attractiveness Stereotype,“ *Psychological Bulletin* 110 (1 1991): 109–128, doi:10.1037/0033-2909.110.1.109.

125 M. G. Efran and E. W. J. Patterson, „The Politics of Appearance“ (Neuverejnená dizertačná práca, 1976).

126 Tamtiež.

127 Thomas Lee Budesheim and Stephen DePaola, „Beauty or the Beast?: The Effects of Appearance, Personality, and Issue Information on Evaluations of Political Candidates,“ *Personality and Social Psychology Bulletin* 20 (4 1994): 339–348, doi:10.1177/0146167294204001.

128 Denise Mack and David Rainey, „Female Applicants’ Grooming and Personnel Selection,“ *Journal of Social Behavior and Personality* 5 (5 1990): 399–407.

129 Daniel S. Hamermesh and Jeff E. Biddle, „Beauty and the Labor Market,“ [Krása a trh práce] *The American Economic Review* 84 (5 1994): 1174–1194.

> Rovnako znepokojujúci výskum naznačuje, že náš súdny proces podobne podlieha vplyvu telesných rozmerov a štruktúry kostí. Vyzerá to, že dobre vyzerajúci ľudia majú väčšiu šancu, že sa k nim právny systém zachová veľmi výhodne (viď Castellow, Wuensch, & Moore, 1991; a Downs & Lyons, 1990, prehľadové štúdie).[130] Napríklad v štúdii z Pennsylvanie (Stewart, 1980)[131] výskumníci hodnotili fyzickú príťažlivosť 74 mužov nezávisle obžalovaných na začiatku súdneho procesu. Keď si omnoho neskôr výskumníci pozreli súdne záznamy výsledkov z týchto prípadov, zistili, že atraktívni muži dostali výrazne ľahšie tresty. V skutočnosti mali príťažliví obžalovaní dvakrát väčšiu šancu vyhnúť sa väzeniu ako nepríťažliví obžalovaní. V inej štúdii – táto sa týkala odškodného prideleného vo fingovanom súdnom procese ublíženia z nedbanlivosti – obvinený, ktorý vyzeral lepšie ako jeho obeť, bol odsúdený zaplatiť v priemere 5 623 dolárov; ale ak bola obeť príťažlivejšia, priemerné odškodné bolo 10 051 dolárov. Navyše, sklon uprednostňovať príťažlivejších vykazovali aj muži aj ženy v porote (Kulka & Kessler, 1978).[132]

Iné pokusy ukázali, že príťažliví ľudia skôr získajú pomoc v núdzi (Benson, Karabenic, & Lerner, 1976)[133] a sú presvedčivejší pri menení názorov obecenstva (Chaiken, 1979).[134]

Tento vplyv príťažlivosti na hodnotenie inteligencie, čestnosti alebo láskavosti je jasným príkladom skreslenia – najmä keď tie druhé posudzujete na základe daného textu – pretože by sme nečakali, že sa naše úsudky o čestnosti a príťažlivosti budú spájať z nejakého legitímneho dôvodu. Na druhej strane, aká časť mojej vnímanej inteligencie je vďaka mojej čestnosti? Aká časť mojej vnímanej čestnosti je vďaka mojej inteligencii? Nájsť pravdu a povedať pravdu nie je vo svojej podstate až tak výrazne odlišné ako vyzerať pekne a vyzerať bystro...

Ale tieto štúdie efektu svätožiary u príťažlivosti by v nás mali vzbudiť podozrenie, že podobný efekt svätožiary môže byť aj u láskavosti alebo inteligencie. Povedzme, že poznáme niekoho, kto vyzerá nielen veľmi inteligentne, ale aj čestne, altruisticky, milo a pokojne. Mali by ste sa zamyslieť, či niektoré z týchto vnímaných vlastnosti nie sú ovplyvnené naším vnímaním iných. Možno je tá osoba naozaj inteligentná, čestná a altruistická, ale nie až taká milá alebo pokojná. Mali by ste sa zamyslieť, ak sa ľudia, ktorých poznáte, delia príliš jasne na anjelov a diablov.

A – viem, že si nemyslíte, že to práve *vy* musíte urobiť, ale možno by ste mali – buďte o trochu skeptickejší pri príťažlivých politických kandidátoch.

* →

104. Skreslenie superhrdinu

Predstavte si, že je tu ťažko ozbrojený sociopat, únosca s rukojemníkmi, ktorý práve odmietol všetky pokusy o vyjednávanie a oznámil svoj úmysel začať zabíjať rukojemníkov. V skutočnom živote kladní hrdinovia zvyčajne nevykopávajú dvere, ak má záporný hrdina rukojemníkov. Ale niekedy – *veľmi*

130 Wilbur A. Castellow, Karl L. Wuensch, and Charles H. Moore, „Effects of Physical Attractiveness of the Plaintiff and Defendant in Sexual Harassment Judgments,“ *Journal of Social Behavior and Personality* 5 (6 1990): 547–562; A. Chris Downs and Phillip M. Lyons, „Natural Observations of the Links Between Attractiveness and Initial Legal Judgments,“ *Personality and Social Psychology Bulletin* 17 (5 1991): 541–547, doi:10.1177/0146167291175009.

131 John E. Stewart, „Defendants' Attractiveness as a Factor in the Outcome of Trials: An Observational Study,“ *Journal of Applied Social Psychology* 10 (4 1980): 348–361, doi:10.1111/j.1559-1816.1980.tb00715.x.

132 Richard A. Kulka and Joan B. Kessler, „Is Justice Really Blind?: The Effect of Litigant Physical Attractiveness on Judicial Judgment,“ *Journal of Applied Social Psychology* 8 (4 1978): 366–381, doi:10.1111/j.1559-1816.1978.tb00790.x.

133 Peter L. Benson, Stuart A. Karabenick, and Richard M. Lerner, „Pretty Pleases: The Effects of Physical Attractiveness, Race, and Sex on Receiving Help,“ *Journal of Experimental Social Psychology* 12 (5 1976): 409–415, doi:10.1016/0022-1031(76)90073-1.

134 Shelly Chaiken, „Communicator Physical Attractiveness and Persuasion,“ *Journal of Personality and Social Psychology* 37 (8 1979): 1387–1397, doi:10.1037/0022-3514.37.8.1387.

→ http://lesswrong.com/lw/lj/the_halo_effect/

zriedkavo, ale predsa – život napodobňuje Hollywood do tej miery, že skutoční kladní hrdinovia musia vylamovať dvere.

Predstavte si dve široko oddelené skutočnosti, dvoch hrdinov, ktorí vtrhnú do miestnosti, aby ako prví čelili lotrovi.

V jednej skutočnosti je hrdina taký silný, že dokáže hádzať autá, môže z nosa vyfukovať ohnivé strely, má rentgenový zrak a jeho koža nielenže *odráža* guľky, ale ich vyparí pri dotyku. Lotor sa opevnil v základnej škole a vzal si ako rukojemníkov vyše dvesto detí; ich rodičia čakajú vonku a plačú.

V druhej skutočnosti je hrdina policajt v New Yorku a rukojemníci sú tri prostitútky, ktoré lotor zobral z ulice.

Zvážte veľmi pozorne túto otázku: Kto je väčší hrdina? A kto s väčšou pravdepodobnosťou bude mať svoj vlastný komix?

Efekt svätožiary je, že vnímanie všetkých kladných vlastnosti koreluje. Profily hodnotené vysoko na škále príťažlivosti sú zároveň hodnotené vyššie na škále talentu, láskavosti, úprimnosti a inteligencie.

A tak komixové postavy, ktoré vyzerajú silné a nezraniteľné, čo sú obe kladné vlastnosti, vyzerajú, že majú aj viac odvahy a hrdinstva. A predsa:

> Aké ťažké to vôbec môže byť správať sa statočne a odvážne, keď ste viacmenej nezraniteľní?
>
> --Adam Warren, Empowered, 1. časť[135]

Nepamätám sa, či som nasledujúce niekde čítal alebo či som to vymyslel sám: Zdá sa, že napríklad *sláva* sa sčítava so všetkými ostatnými osobnostnými črtami. Vezmime si Gándhího. Bol Gándhí *najväčším altruistom* 20. storočia, alebo iba *najznámejším* altruistom? Gándhí čelil policajtom s obuškami a vojakom s puškami. Lenže Gándhí bol celebrita a to ho chránilo. Čo zvyšní účastníci pochodu, ktorí čelili obuškom a puškám, hoci by neboli žiadne medzinárodné správy, keby skončili v nemocnici alebo zastrelení?

Čo si Gándhí myslel o tom, že dostal titulky, stal sa celebritou, získal slávu a miesto v dejinách, *stal sa archetypom* nenásilného odporu, keď riskoval menej než hocikto z tých, ktorí pochodovali s ním? Ako sa cítil, keď niektorý z týchto anonymných hrdinov prišiel k nemu s rozžiarenými očami a povedal Gándhímu, aký je úžasný. Videl vôbec Gándhí niekedy svoj svet z tohto uhla pohľadu? Neviem; nie som Gándhí.

Toto nie je v žiadnom zmysle kritika Gándhího. Pointa nenásilného odporu nie je predvádzať svoju odvahu. To sa dá urobiť omnoho ľahšie, keď sa spustíte z niagarského vodopádu v sude. Gándhí nemal ma výber ohľadom toho, že bol trochu, aj keď nie úplne, chránený ako celebrita. A Gándhího činy si vyžadovali odvahu – nie toľko odvahy ako pochodovať anonymne, ale aj tak dosť odvahy.

Skreslenie, na ktoré chcem ukázať je, že Gándhího body za slávu sa v našom vnímaní *pripočítajú* k jeho oprávnene získaným bodom za altruizmus. Keď myslíte na nenásilie, myslíte na Gándhího – nie na nejakú anonymnú demonštrantku v jednom z Gándhího pochodov, ktorá čelila obuškom a puškám, dostala bitku, musela ísť do nemocnice, celý zvyšok svojho života krívala, *a nikto si nikdy nespomenie na jej meno.*

Podobne, čo je viac – riskovať svoj život, aby ste zachránili dvesto detí alebo riskovať svoj život, aby ste zachránili troch dospelých?

Odpoveď závisí od toho, čo myslíme slovom *viac*. Ak si niekedy budete musieť *vybrať* medzi záchranou troch dospelých a záchranou dvesto detí, vyberte si tú druhú možnosť. „Kto zachráni jeden život, je to akoby zachránil celý svet“ môže byť pekný signál na potlesk, ale je to hrozná morálna rada, ak si máte vybrať jedno alebo druhé. Ak teda myslíte „viac“ v zmysle „Ktoré je dôležitejšie?“ alebo „Ktoré je želanejší výsledok?“ alebo „Ktoré by som si mal vybrať, keby som si musel vybrať jedno alebo druhé?“ potom je viac zachrániť dvesto než troch.

135 Adam Warren, *Empowered*, vol. 1 (Dark Horse Books, 2007).

Ale ak sa pýtate v zmysle, čo je väčšia prejavená cnosť, potom niekto, kto je ochotný riskovať svoj život, aby zachránil iba tri životy, prejavil viac odvahy než niekto, kto je ochotný riskovať svoj život, aby zachránil dvesto, ale nie tri.

To neznamená, že si môžete vedome vybrať ísť riskovať svoj život kvôli záchrane troch dospelých a nechať dvesto školákov osudu, pretože chcete prejaviť viac cnosti. Niekto, kto riskuje svoj život, *pretože chce byť cnostný*, prejavil omnoho menej cnosti ako niekto, kto riskuje svoj život, *pretože chce zachrániť druhých*. Niekto, kto si vyberie zachrániť tri životy namiesto dvesto životov, pretože si myslí, že tým prejaví viac cnosti, je tak sebecky fascinovaný svojou vlastnou „veľkosťou“, akoby spáchal morálny ekvivalent zabitia.

Je to jeden z tých scenárov *wu wei*: Nemôžete prejaviť cnosť snahou prejaviť cnosť. Ak dostanete na výber medzi bezpečnou metódou ako zachrániť svet, ktorá nevyžaduje žiadne osobné obete ani nepohodlie a metódou, pri ktorej riskujete svoj život a musíte podstúpiť mnohé príkoria, nemôžete sa stať hrdinom tým, že si úmyselne vyberiete druhú možnosť. Nie je nič hrdinské na tom chcieť byť hrdinom. Je to stratený cieľ.

Skutočne cnostní ľudia, ktorí sa naozaj snažia zachrániť životy, namiesto snahy prejaviť cnosť, budú vytrvalo hľadať, ako zachrániť viac životov s menšou námahou, čo znamená, že sa prejaví menej ich cnosti. Môže to byť mätúce, ale nie je to rozpor.

Nemôžeme si však vždy vybrať, že budeme nezraniteľní guľkami. Potom, čo sme spravili, čo bolo v našich silách na zníženie rizika a zvýšenie rozsahu, všetko *zostávajúce* hrdinstvo sa dobre a pravdivo prejaví.

Policajt, ktorý nasadzuje svoj život v bojovej línii bez superschopností, bez rentgenového pohľadu, bez super sily, bez schopnosti lietať, a najmä bez odolnosti voči guľkám, prejavuje omnoho viac cnosti ako Superman – ktorý je *iba superhrdina*.

*→

105. Iba spasitelia

Včera som diskutoval, ako efekt svätožiary, ktorý spôsobuje, že ľudia vidia všetky kladné vlastnosti ako korelované – napríklad príťažlivejší jednotlivci sú zároveň vnímaní ako láskavejší, úprimnejší a inteligentnejší – spôsobuje, že viac obdivujeme hrdinov, ak sú super silní a odolní voči guľkám. Hoci logicky si vyžaduje omnoho viac odvahy byť hrdinom, ak *nie* ste odolní voči guľkám. Ďalej, že prejavuje viac cnosti konať odvážne, aby ste zachránili jeden život, než aby ste zachránili svet. (Hoci ak máte na výber jedno alebo druhé, samozrejme by ste mali zachrániť svet.)

„Policajt, ktorý nasadí svoj život v bojovej línii bez superschopností,“ povedal som, „prejaví omnoho viac cnosti než Superman, ktorý je *iba superhrdina*.“

Ale buďme konkrétnejší.

John Perry bol policajt v New Yorku, ktorý bol zároveň Extropián a transhumanista, kde som ho spoznal po mene. John Perry mal o chvíľu ísť do dôchodku a založiť si právnu prax, keď sa dozvedel, že lietadlo narazilo do World Trade Center. Zomrel pri zrútení severnej veže. Nepoznal som Johna Perryho osobne, takže toto nemôžem potvrdiť z prvej ruky; ale veľmi málo Extropiánov verilo v Boha a predpokladám, že aj John Perry bol ateista.

Čo znamená, že Perry vedel, že riskuje svoju samotnú existenciu, každý týždeň v práci. Nie je to tak, že by ako väčšina ľudí v dejinách vedel, že má na výber iba ako zomrie, a tak si vybral zmysluplnú možnosť – Perry bol totiž transhumanista; mal skutočnú nádej. A Perry tam šiel a napriek tomu riskoval svoj život. Nie preto, že by čakal nejakú nadprirodzenú odmenu. Nie preto, že by očakával vôbec niečo po svojej prípadnej smrti. Ale pretože tam boli v nebezpečenstve iní ľudia a oni tiež nemali nehmotné duše, a jeho nádej na život nemala väčšiu cenu než ich.

→ http://lesswrong.com/lw/lk/superhero_bias/

Nepoznal som Johna Perryho. Neviem, či vnímal svet týmto spôsobom. Ale fakt, že ateista a transhumanista môže stále byť policajtom, môže vbehnúť do čakárne v horiacej budove, hovorí o ľudskom duchu viac než všetci mučeníci, ktorí kedy dúfali v nebo.

Takže toto je jeden konkrétny policajt...

...a teraz ten superhrdina.

Ako ten príbeh rozprávajú kresťania, Ježiš Kristus vedel chodiť po vode, tíšiť búrky, vyháňať slovom démonov. Musel to byť pohodlný život: Hrozí hlad? Nakopírujem si trochu chleba. Nepáči sa mi strom? Prekľajem ho. Rimania robia problém? Pošlem na nich Otca. Nakoniec tento čarovný život skončil a Ježiš sa dobrovoľne ponúkol na ukrižovanie. Byť priklincovaný na kríž nie je pohodlný spôsob smrti. Ale ako ten príbeh rozprávajú kresťania, Ježiš to urobil s vedomím, že o tri dni neskôr ožije a potom pôjde do Neba. Čo bola tá hrozba, ktorá pohla Ježiša čeliť tomuto dočasnému utrpeniu nasledovanému večnosťou v Nebi? Bol to život jedného človeka? Bola to skorumpovanosť cirkvi v Judei alebo útlak Ríma? Nie: ako ten príbeh rozprávajú kresťania, v ohrození bol večný osud každého človeka, kým sa Ježiš nechal dočasne priklincovať na kríž.

Ja nechcem až takto odsudzovať človeka, ktorý nie je naozaj taký vinný. Čo ak Ježiš – nie, vyslovujme jeho meno správne: Jeišu – čo ak Jeišu z Nazaretu nikdy nechodil po vode a *predsa* vzdoroval cirkvi v Judei podporovanej mocou Ríma?

Nezaslúžilo by si toto viac úcty než na akú si môže nárokovať Ježiš Kristus, ktorý bol iba spasiteľ?

Žiaľ, akosi sa zdá väčším hrdinom ten, kto má oceľovú kožu a božskú moc. Akosi sa zdá, že prejavuje viac cnosti zomrieť dočasne pre záchranu celého sveta než zomrieť natrvalo v konflikte so skorumpovanou cirkvou. Zdá sa to také *všedné*, akoby mnoho ďalších ľudí v dejinách urobilo to isté.

Pohodlne usadení dvetisíc rokov v budúcnosti môžeme Jeišua zahrnúť všetkou možnou kritikou, ale Jeišu robil to, o čom veril, že je správne, postavil sa proti cirkvi, o ktorej veril, že je skazená, a zomrel za to. Bez výhody spätného pohľadu od neho sotva možno očakávať, že by predpovedal skutočný dopad svojho života na svet. V porovnaní s inými prorokmi svojej doby bol pravdepodobne pomerne úprimnejší, pomerne menej násilný, a pomerne odvážnejší. Ak neberieme do úvahy neplánované dôsledky, najhoršia vec, ktorú možno o Jeišuovi povedať je, že iní to v dejinách urobili lepšie. (Napadajú mi Epikuros, Buddha, Markus Aurelius.) Jeišu zomrel navždy a – z jedného pohľadu – to urobil kvôli úprimnosti. Pätnásť storočí pred vedou náboženská úprimnosť nebola oxymoron.

Ako povedal Sam Harris:[136]

> Nestačí, že Ježiš bol človek, ktorý sa premenil natoľko, že Kázeň na hore mohla byť spoveďou jeho srdca. Musel byť ešte aj Synom Boha, narodený z panny a predurčený vrátiť na zem oblaky slávy. Dôsledkom takejto dogmy je umiestnenie Ježišovho príkladu navždy mimo náš dosah. Jeho učenie prestane byť množinou empirických tvrdení o vzťahu medzi etikou a duchovným vhľadom a namiesto toho sa stane samoúčelnou a trochu morbídnou rozprávkou. Podľa kresťanských dogiem stať sa ako Ježiš je nemožné. Človek môže iba vymenovávať svoje hriechy, veriť v neuveriteľné a čakať na koniec sveta.

Vážne pochybujem, že Jeišu niekedy hovoril Kázeň na hore. Napriek tomu si Jeišu zaslúži úctu. Zaslúži si viac úcty než by mu kresťania ponechali.

Ale keďže Jeišu pravdepodobne očakával, že jeho duša prežije, nezaslúži si viac úcty než John Perry.

*→

106. Afektívne špirály smrti

Mnoho, mnoho, mnoho je chýb v ľudskom uvažovaní, ktoré nás vedú preceňovať ako dobre naša obľúbená teória vysvetľuje fakty. Chemická teória flogistonu mohla vysvetliť takmer všetko, pokiaľ

136 Sam Harris, *The End of Faith: Religion, Terror, and the Future of Reason*, (WW Norton & Company, 2005).
→ http://lesswrong.com/lw/ll/mere_messiahs/

nemusela niečo predpovedať vopred. A na vysvetlenie čím väčšieho množstva javov ste použili svoju obľúbenú teóriu, tým pravdivejšia vaša obľúbená teória vyzerá – nebola vari potvrdená tými mnohými pozorovaniami? Čím pravdivejšia vyzerá táto teória, tým skôr budete spochybňovať indície, ktoré jej protirečia. Čím všeobecnejšia sa zdá táto obľúbená teória, na tým viac vysvetlení ju budete používať.

Ak poznáte niekoho, kto verí, že Belgicko tajne ovláda americký bankový systém, alebo že môžu použiť neviditeľnú modrú duchovnú silu na nájdenie voľného parkovacieho miesta, takto nejako to asi začalo.

(Zostaňte v strehu a uvidíte toľko vecí, ktoré budú vyzerať, že potvrdzujú túto teóriu...)

Tento cyklus pozitívnej spätnej väzby dôverčivosti a potvrdenia je naozaj obávaný a zodpovedný za mnoho chýb, vo vede aj v bežnom živote.

Ale to nie je nič v porovnaní so špirálou smrti, ktorá začne nábojom pozitívneho citu – myšlienkou, ktorá *vyzerá naozaj dobre*.

Nový politický systém, ktorý môže zachrániť svet. Veľký vodca, silný a vznešený a múdry. Zázračný nápoj, ktorý vylieči boľavé brucho aj rakovinu.

Sakra, prečo nie všetky tri? Veľká úloha si vyžaduje veľkého vodcu. Veľký vodca by mal dokázať namiešať pár čarovných nápojov.

Efekt svätožiary je, že každá vnímaná kladná vlastnosť (napríklad príťažlivosť alebo sila) zvyšuje vnem ľubovoľnej inej kladnej vlastnosti (napríklad inteligencie alebo odvahy). Dokonca aj keď to nedáva zmysel, alebo je to proti zmyslu.

Kladné vlastnosti zvyšujú vnímanie každej ďalšej kladnej vlastnosti? To znie podobne ako keď rozpadajúci sa atóm uránu vystrelí neutróny, ktoré rozbijú ďalšie atómy uránu.

Slabý kladný cit je podkritický; nezačne sa vymykať spod kontroly. Príťažlivá osoba vyzerá úprimnejšia, čo ju povedzme robí ešte trochu atraktívnejšou; ale účinný násobiteľ neutrónov je menší ako 1. Metaforicky povedané. Rezonancia trochu zamieša veci ale potom odumrie.

Keď sa intenzívny kladný cit pripojí k Veľkej Veci, rezonancia zasiahne všetko. Veriaci komunista vidí Marxovu múdrosť v každom hamburgeri kúpenom v McDonalde; v každom povýšení, ktoré mu zamietli a ktoré by dostal v skutočnom robotníckom raji; v každých voľbách, ktoré nedopadnú podľa jeho vkusu; v každom „nesprávne orientovanom“ novinovom článku. Vždy keď použije Veľkú Myšlienku na interpretovanie ďalšej udalosti, Veľká Myšlienka vyzerá byť ešte potvrdenejšia. Je to dobrý pocit – pozitívne posilnenie – a samozrejme, keď máme z niečoho dobrý pocit, žiaľ, o to viac tomu *chceme* veriť.

Keď máte z Veľkej Veci taký dobrý pocit, že *idete hľadať* ďalšie príležitosti ako mať z Veľkej Veci ešte lepší pocit, používate ju na vysvetlenie nových udalostí každý deň, rezonancia kladného citu je ako miestnosť plná pascí na myši, do ktorej nasypete pingpongové loptičky.

Môžete to nazvať „šťastný atraktor“, „príliš pozitívna spätná väzba“, „uzavretá slučka chvály“ alebo „zábavné noviny“. Ja osobne dávam prednosť slovnému spojeniu „afektívna špirála smrti“.

Nabudúce: Ako odolať afektívnej špirále smrti. (Nápoveda: Nie tak, že odmietnete kedykoľvek čokoľvek obdivovať, ani že budete držať obdivované veci v malých bezpečných ohraničených magistériách.)

*→

107. Odolajte šťastnej špirále smrti

Kde bolo, tam bolo, bol jeden človek, ktorý bol presvedčený, že objavil Veľkú Myšlienku. Skutočne, ako ten človek rozmýšľal o tejto Veľkej Myšlienke viac a viac, uvedomoval si, že je to nielen *nejaká* veľká myšlienka, ale tá *najúžasnejšia myšlienka všetkých čias*. Táto Veľká Myšlienka odhalí tajomstvá vesmíru, nahradí autoritu skorumpovaného systému plného chýb, dá svojim nositeľom takmer magickú moc, nasýti hladných, uzdraví chorých, urobí z celého sveta lepšia miesto, atď. atď. atď.

→ http://lesswrong.com/lw/lm/affective_death_spirals/

Ten človek bol Francis Bacon a jeho Veľká Myšlienka bola vedecká metóda a bol to jediný fanatik v dejinách, ktorý si pripisoval takýto stupeň úžitku pre ľudstvo, a ukázalo sa, že má úplnú pravdu.

(Bacon samozrejme nevynašiel vedu sám, ale prispel a možno bol prvý, kto si uvedomil jej moc.)

Toto je problém rozhodnutia, že nikdy nebudete nič tak veľmi obdivovať: Niektoré myšlienky naozaj *sú* také dobré. Aj keď nikto *nesplnil* sľuby odvážnejšie než tie Baconove; prinajmenšom zatiaľ.

Ako však potom môžeme odolať šťastnej špirále smrti voči Vede samotnej? Šťastná špirála smrti začína vtedy, keď veríte, že niečo je *také* úžasné, že vás efekt svätožiary vedie k hľadaniu *ďalších* a *ďalších* pekných vecí, ktoré o tom môžete povedať, vďaka čomu to vidíte ako *ešte* úžasnejšie a tak ďalej, po špirále smerom do priepasti. Čo ak je Veda *naozaj* taká úžasná, že nemôžeme pripustiť jej skutočnú slávu a zachovať si pritom príčetnosť? To znie ako pekná vec, nie? *Ach nie, už to začína, utekajteee...*

Ak vyberiete štandardnú uloženú hlbokú múdrosť na *nepreháňanie obdivu vedy*, nájdete myšlienky ako: „Veda nám dala klimatizáciu, ale aj vodíkovú bombu“ alebo „Veda nám môže povedať o hviezdach a biológii, ale nevie dokázať ani vyvrátiť draka v mojej garáži“. Lenže ľudia, od ktorých *pochádzajú* takéto myšlienky sa *nepokúšali* odolať šťastnej špirále smrti. Neznepokojovali sa, že sa ich vlastný obdiv vedy vymkne spod kontroly. Pravdepodobne sa im nepáčilo niečo, čo veda povedala o ich obľúbenom názore a hľadali, ako podkopať jej autoritu.

Štandardné negatívne myšlienky, ktoré sa hovoria o vede, pravdepodobne nebudú pôsobiť na niekoho, kto naozaj cíti nadšenie z vedy – to nie je cieľová skupina. Musíme teda namiesto toho hľadať iné zlé veci, ktoré možno povedať.

Ak však selektívne hľadáte niečo zlé, čo môžete povedať o vede – hoci aj v snahe odolať šťastnej špirále smrti – neusvedčujete automaticky sami seba z racionalizovania? Prečo by ste mali venovať pozornosť svojim myšlienkach, keď viete, že sa snažíte sami seba manipulovať?

Vo všeobecnosti som skeptický voči ľuďom, ktorí tvrdia, že jedno skreslenie možno použiť na vyváženie druhého. Znie mi to ako automechanik, ktorý povie, že na pravom stierači máte pokazený motor, ale namiesto opravy vám jednoducho zlomí ľavý stierač, aby sa tým veci vyrovnali. Takýto druh chytrosti vedie k strieľaniu vlastných gólov. Nech je riešením čokoľvek, mali by sme v rámci neho veriť pravdivým veciam, nie veriť, že veríme veciam, o ktorých veríme, že sú nepravdivé.

Dokážete zabrániť šťastnej špirále smrti tým, že svoj obdiv vedy obmedzíte na úzku oblasť? Časť šťastnej špirály smrti je vidieť Veľkú Myšlienku všade – myslieť na to, ako by komunizmus vyliečil rakovinu, keby len dostal príležitosť. Pravdepodobne najspoľahlivejším znakom vodcu sekty je, že predstiera odbornosť nie v jednej oblasti, ani v skupine súvisiacich oblastí, ale vo *všetkom*. Vodca vie, čo by členovia sekty mali jesť, čo si obliekať, čím sa živiť; s kým by mali mať sex; na aké umenie by sa mali pozerať; akú hudbu by mali počúvať...

Nanešťastie pre tento plán, väčšina ľudí biedne zlyháva, keď sa snažia opísať tú peknú malú krabičku, vo vnútri ktorej by veda mala zostať. Zvyčajný trik: „Pozri, veda nedokáže vyliečiť rakovinu“ tu nezaberie. „Veda nedokáže nič povedať o láske rodiča k dieťaťu“ - pardon, to je jednoducho nepravda. Ak sa pokúsite odseknúť vedu napríklad od rodičovskej lásky, tým nepopierate iba kognitívnu vedu a evolučnú psychológiu. Popierate aj, že Martine Rothblatt založila United Therapeutics s cieľom nájsť liek na pľúcnu hypertenziu svojej dcéry. (Dodávam, že úspešne.) Veda je legitímne spojená, jedným či druhým spôsobom, s takmer každou dôležitou stránkou ľudskej existencie.

No dobre, tak čo je príkladom *nepravdivého* pekného tvrdenia, ktoré nemôžete povedať o vede?

Podľa môjho skromného názoru jedno nepravdivé tvrdenie je, že veda je taká úžasná, že vedci by sa nemali ani len snažiť prevziať etickú zodpovednosť za svoju prácu, pretože to automaticky skončí dobre. Toto tvrdenie podľa mňa nerozumie podstate procesu, ktorým veda pomáha ľudstvu. Vedci sú ľudia, majú prosociálny záujem tak ako väčšina ľudí, a toto je prinajmenšom *časť* toho, prečo veda v konečnom dôsledku robí viac dobra než zla.

Ale tento bod očividne nie je mimo debaty. Tu je teda jednoduchšie nepravdivé pekné tvrdenie: „Pacienta s rakovinou možno uzdraviť púhym publikovaním dostatočného množstva článkov

v odborných časopisoch.“ Alebo: „Sociopati sa môžu stať celkom normálni, ak si dajú záväzok nikdy neveriť ničomu bez indície replikovaných experimentov s $p < 0,05$.“

Viere v takéto tvrdenia sa nevyhýbame tým, že si stanovíme citovú hranicu a rozhodneme sa, že veda je iba trochu pekná. Ani hľadaním dôvodov veriť, že uverejňovanie odborných článkov *spôsobuje* rakovinu. Ani vierou, že veda nedokáže o rakovine povedať nič dobré ani zlé.

Namiesto toho, ak viete dostatočne podrobne ako veda funguje, potom viete, že hoci je možné, aby „veda vyliečila rakovinu“, pacient s rakovinou píšuci články do odborných časopisov nezažije zázračnú remisiu. Táto *konkrétna* reťaz príčin a následkov nebude fungovať.

Šťastná špirála smrti je emocionálnym problémom iba kvôli problému vnímania, efektu svätožiary, ktorý nás robí ochotnejšími prijať ďalšie pozitívne tvrdenia, keď už sme prijali úvodné pozitívne tvrdenie. Tohto efektu sa nezbavíme tým, že si to jednoducho budeme želať; pravdepodobne nás vždy bude trochu ovplyvňovať. Ale dokážeme spomaliť, zastaviť sa, zvážiť každé dodatočné pekné tvrdenie ako dodatočnú priťažujúcu podrobnosť, a zamerať sa na konkrétne časti tvrdenia bez ohľadu na jeho pozitívnosť.

Čo ak konkrétne pekné tvrdenie „nemožno vyvrátiť“, ale existujú argumenty „za aj proti“ tomuto tvrdeniu? V skutočnosti si na tieto slová treba dávať pozor vo všeobecnosti, pretože toto ľudia často hovoria, keď si precvičujú indície alebo sa vyhýbajú naozaj slabým miestam. Pri nebezpečenstve šťastných špirál smrti dáva zmysel skúsiť sa vyhýbať radosti z *neuzavretých* tvrdení – aby sa nestali zdrojom ďalších ešte pozitívnejších citov voči niečomu, čo už máte radi.

Šťastná špirála smrti je *veľký* emocionálny problém iba kvôli príliš pozitívnej spätnej väzbe, kvôli schopnosti procesu prekročiť kritické množstvo. Možno nebudete vedieť úplne odstrániť efekt svätožiary, ale môžete použiť dosť kritického myslenia, aby ste svätožiaru udržali pod kritickým množstvom – aby ste zabezpečili, že rezonancia vyhasne namiesto vybuchnutia.

Mohli by ste priam povedať, že celý problém začína, keď sa ľuďom nechce kriticky skúmať každú priťažujúcu podrobnosť navyše – vyžadovať dostatočnú indíciu na vyváženie zložitosti, hľadať chyby, nielen podporu, používať zvedavosť – akonáhle prijali nejaký základný predpoklad. Bez klamu konjunkcie by stále mohol byť efekt svätožiary, ale nebola by šťastná špirála smrti.

Ešte aj pri tých najkrajších Pekných Veciach v známom vesmíre by dokonalý racionalista, ktorý by presne vyžadoval potrebné indície na každé ďalšie (pozitívne) tvrdenie, nezažil afektívnu rezonanciu. Vy toto nedokážete, ale môžete sa držať dosť blízko rozumnosti, aby sa vaše šťastie nevymklo spod kontroly.

Naozaj nebezpečné prípady sú tie, kde *hocijaká kritika nejakého pozitívneho tvrdenia o Veľkej Veci vyzerá ako zlá alebo spoločensky neprijateľná.* Argumenty sú vojaci, každé pozitívne tvrdenie je vojak na našej strane, bodať vlastných vojakov do chrbta je vlastizrada. Potom sa reťazová reakcia stane *nadkritickou*. Viac o tomto zajtra.

Stuart Armstrong ponúka súvisiacu radu:

> Rozdeľ svoju Veľkú Vec na menšie nezávislé myšlienky a *ber ich ako nezávislé.*
>
> Napríklad marxista by mohol rozdeliť Marxovu Veľkú Vec na teóriu hodnoty práce, teóriu politických vzťahov medzi triedami, teóriu mzdy, teóriu konečného politického stavu ľudstva. Potom by mal každé z toho vyhodnotiť nezávisle a ich pravdivosť alebo nepravdivosť by nemala presakovať do zvyšných. Ak dokážeme urobiť toto, mali by sme byť v bezpečí pred špirálou, keďže každá teória je príliš úzka na to, aby začala svoju vlastnú špirálu.

Metaforicky je to ako držať podkritické množstvá plutónia oddelene od seba. Tri Veľké Myšlienky vás omnoho ťažšie doženú do šialenstva než jedna Veľká Myšlienka. Armstrongova rada zároveň propaguje konkrétnosť. Akonáhle niekto povie: „Publikovanie dostatočného množstva odborných článkov môže vyliečiť vašu rakovinu“, opýtajte sa: „Je toto dôsledkom experimentálnej metódy, a ak áno, v ktorom kroku experimentálneho procesu sa tá rakovina vylieči? Alebo je to dôsledok vedy ako spoločenského procesu, a ak áno, závisí to od toho, či jednotlivý vedec chce vyliečiť rakovinu, alebo môže jednať v sebeckom záujme?“ Dúfajme, že toto povedie preč od dobrých či zlých pocitov, smerom k uvedomeniu si zmätenia a nedostatku podpory.

Aby som to zhrnul, šťastnej špirále smrti sa *vyhnete*:

- Rozdelením Veľkej Myšlienky na časti;
- Považovaním každého dodatočného detailu za priťažujúci;
- Rozmýšľaním o konkrétnych častiach kauzálnej reťaze namiesto o dobrých či zlých pocitoch;
- Neprecvičovaním si indícií;
- Nepridávaním šťastia z tvrdení, o ktorých: „nemôžete *dokázať*, že sú nesprávne“;

ale *nie*:

- Odmietaním čokoľvek príliš obdivovať;
- Podniknutím skresleného hľadania zlých stránok, dokiaľ sa opäť nezačnete cítiť nešťastne;
- Násilným strčením myšlienky do bezpečnej krabice.

*→

108. Nekritická nadkritickosť

Z času na čas vidíte, ako sa ľudia hádajú, či je ateizmus „náboženstvo“. Ako som naznačil v článku Cieľ a pragmatizmus, hádanie sa o význame slova takmer vždy znamená, že ste stratili z očí pôvodnú otázku. Ako vôbec môže takáto hádka začať?

Ateista reční, obviňuje „náboženstvo“ za inkvizíciu, krížové výpravy a rôzne konflikty s islamom alebo vnútri neho. Veriaci môže odpovedať: „Ale ateizmus je tiež náboženstvo, pretože aj vy máte vieru ohľadom Boha; veríte, že Boh neexistuje.“ Potom ateista povie: „Ak ateizmus je náboženstvo, potom nezbieranie známok je hobby,“ a začne hádka.

Prípadne možno namietať: „Ale rovnako veľké hrôzy spôsobil Stalin, ktorý bol ateista a potláčal cirkvi v mene ateizmu; preto je nesprávne obviňovať z násilia náboženstvo.“ Na to môže mať ateista chuť povedať: „Stalinovým náboženstvom bol komunizmus.“ Na to veriaci odpovie: „Ak komunizmus je náboženstvo, potom fandom Star Wars je vláda“ a začne hádka.

Mal by „náboženský“ názor byť definovaný ako definitívny názor na existenciu aspoň jedného Boha, napríklad priradenie pravdepodobnosti nižšej ako 10 % alebo vyššej ako 90 % existencii Zeusa? Alebo by „náboženský“ názor mať byť definovaný ako pozitívny názor, povedzme pravdepodobnosť vyššia ako 90 % v prospech existencie aspoň jedného Boha? V tom prvom prípade bol Stalin „nábožný“; v tom druhom prípade bol Stalin „nie nábožný“.

Ale práve toto je nesprávny spôsob, ako pozerať na problém. To, čo naozaj chcete vedieť – o čom tá hádka bola na začiatku – je prečo boli v niektorých bodoch ľudských dejín veľké skupiny ľudí zabíjané a mučené, napohľad v mene nejakej myšlienky. Predefinovanie slova nezmení historické fakty jedným ani druhým smerom.

Komunizmus bola zložitá katastrofa a možno neexistuje jeden konkrétny *dôvod* pre to celé, žiadna jedna kritická linka v reťazi kauzality. Ale ak by som mal naznačiť základnú chybu, bolo by to... nuž, nech to Boh povie za mňa:

> Ak tvoj brat, syn tvojho otca alebo tvojej matky, alebo tvoj syn alebo dcéra, alebo tvoja manželka, alebo tvoj najlepší priateľ, sa ťa pokúsi tajne zviesť slovami: Poďme a slúžme iným bohom, neznámym tebe a tvojim predkom, bohom okolitých národov, či už blízkych alebo vzdialených, kdekoľvek na svete, nesmieš súhlasiť, **nesmieš ho počúvať**; nesmieš mu prejaviť ľútosť, nesmieš ho ušetriť ani skryť jeho vinu. Nie, **musíš ho zabiť**, tvoja ruka musí prvý úder vedúci k jeho smrti a ruky ostatných ľudí musia nasledovať. Musíš ho ukameňovať na smrť, pretože sa ťa pokúsil odkloniť od Jahveho, tvojho Boha.
>
> --Piata kniha Mojžišova 13:7-11, pridané zvýraznenie

→ http://lesswrong.com/lw/ln/resist_the_happy_death_spiral/

Toto bolo aj pravidlo, ktoré Stalin určil pre komunizmus a Hitler pre nacizmus: ak ti tvoj brat skúša povedať, prečo sa Marx mýli, ak ti tvoj syn skúša povedať, že židia neplánujú svetovládu, nediskutuj s ním ani nepredkladaj svoje indície; nerob replikovateľné pokusy ani neskúmaj históriu; ale okamžite ho udaj tajnej polícii.

Včera som naznačil, že jedným z kľúčov odolávania afektívnej špirále smrti je princíp „priťažujúcich podrobností“ - len si *pamätať*, že treba spochybňovať konkrétne podrobnosti každého dodatočného pekného tvrdenia o Veľkej Myšlienke. (To nie je triviálna rada. Ľudia na toto často zabúdajú, keď počúvajú, ako im futurista vykresľuje úžasne podrobný obraz zázrakov budúcnosti, a tobôž keď myslia na svoju vlastnú najobľúbenejšiu myšlienku.) Toto by nás nezbavilo efektu svätožiary, ale snáď by to obmedzilo rezonanciu pod kritickú hladinu, takže by jedno pekne znejúce tvrdenie vyvolalo v priemere menej ako 1,0 ďalšieho pekne znejúceho tvrdenia.

Pravým opakom tejto rady, ktorý posúva efekt svätožiary *nad* kritickú hladinu, je keď ľuďom pripadá nesprávne argumentovať proti *hocijakému* pozitívnemu tvrdeniu o Veľkej Myšlienke. Politika zabíja myslenie. Argumenty sú vojaci. Akonáhle viete, na ktorej ste strane, musíte podporovať všetky tvrdenia v jej prospech a argumentovať proti všetkým tvrdeniam v jej neprospech. Inak by to bolo ako pomáhať nepriateľovi alebo bodať vlastných vojakov do chrbta.

Ak...

- ...máte pocit, že protirečiť niekomu, kto urobí nesprávne pekné tvrdenie v prospech evolúcie, by znamenalo pomáhať kreacionistom;
- ...máte pocit, že dostávate duchovný kladný bod za každú peknú vec, ktorú poviete o Bohu a že argumentovať proti nej by ohrozilo váš vzťah s Bohom;
- ...ak máte silný pocit, že zvyšní ľudia v miestnosti by vás nemali radi za to, že „nepodporujete naše vojská“, keby ste argumentovali proti najnovšej vojne;
- ...povedanie hocičoho proti komunizmu spôsobí, že vás ukameňujú, pardon, zastrelia;

...potom je afektívna špirála smrti už nadkritická. Už je to Super Šťastná Špirála Smrti.

Teda nie náboženstvo ako také je kľúčovou kategóriou voči našej pôvodnej otázke: „Čo spôsobuje to zabíjanie?“ Najlepšie rozlíšenie, aké som počul medzi „nadprirodzeným“ a „naturalistickým“ svetonázorom je, že nadprirodzený svetonázor tvrdí existenciu ontologicky základných myšlienkových podstát, ako sú duchovia, zatiaľ čo naturalistický svetonázor redukuje myšlienkové javy na nemyšlienkové časti. Sústrediť sa na toto ako na zdroj problémov znamená akceptovať výnimočnosť náboženstva. Nadprirodzené tvrdenia sa oplatí rozlišovať, pretože sa vždy ukážu ako nesprávne z celkom zásadných dôvodov. Ale aj tak to je iba jeden druh chyby.

Afektívna špirála smrti sa môže vytvoriť okolo nadprirodzeného názoru; najmä okolo monoteizmu, ktorého vrcholom je Super Šťastný Činiteľ, definovaný v prvom rade tým, že súhlasíme s každým pekným názorom o ňom; najmä ak komplex mémov narastie na dostatočne sofistikovaný, aby tvrdil, že za nevieru bude nasledovať nadprirodzený trest. Špirála smrti však môže vzniknúť aj okolo politickej inovácie, charizmatického vodcu, viery v osud rasy, alebo ekonomickej hypotézy. Lekcia z dejín znie, že afektívne špirály sú nebezpečné bez ohľadu na to, či obsahujú alebo neobsahujú nadprirodzeno. Náboženstvo nie je dosť špeciálne ako typ chyby, aby bolo kľúčom k tomuto problému.

Sam Harris sa dostal bližšie, keď ukázal obviňujúcim prstom na *vieru*. Ak neukladáte primerané bremeno dôkazu každému jednému dodatočnému peknému tvrdeniu, afektívna rezonancia začne *veľmi* ľahko. Pozrite sa na chudákov z new age. Kresťanstvo si vyvinulo ochrany proti kritike, argumentujúc za zázraky viery; new age kultúrne zdedil uloženú myšlienku, že viera je dobrá, ale chýba mu vylučujúce sväté písmo kresťanstva na bránenie sa konkurujúcim mémom. Priaznivci new age končia v šťastných špirálach smrti okolo hviezd, stromov, magnetov, diét, kúziel, jednorožcov...

Afektívne špirály smrti sa však stávajú omnoho vražednejšie, keď sa kritika stane hriechom, prešľapom, alebo zločinom. Na tomto svete sú veci, ktoré sú hodné veľkej chvály a nemôžete *jednoznačne* povedať, že chvála za istú hranicu je zakázaná. Ale *nikdy* neexistuje Myšlienka natoľko pravdivá, že je nesprávne kritizovať akýkoľvek argument, ktorý ju podporuje. Nikdy. Nikdy nikdy nikdy. *Toto* je jednoznačné. Prevažná väčšina možných názorov v netriviálnom priestore odpovedí je nesprávna,

a podobne aj prevažná väčšina možných *podporujúcich argumentov* pre pravdivý názor je nesprávna, a toto nezmení ani tá najšťastnejšia myšlienka.

A je trojnásobne ultra zakázané reagovať na kritiku násilím. Je iba pár prikázaní v ľudskom umení rozumnosti, ktoré nemajú žiadne „ak“, „a“, „alebo“ a iné únikové doložky. Toto je jedno z nich. Zlý argument má dostať protiargument. Nemá dostať guľku. Nikdy. Nikdy nikdy nikdy.

*→

109. Ochladzovanie skupinových názorov vyparovaním

Prvé štúdie siekt s prekvapením zistili, že keď sekta utrpí väčší šok – keď sa nenaplnilo proroctvo, keď sa odhalilo morálne zlyhanie zakladateľa – často z toho vyjde silnejšia než predtým, so zvýšenou vierou a fanatizmom. Svedkovia Jehovovi určili Armageddon na rok 1975 na základe biblických výpočtov; rok 1975 prišiel a odišiel. Sekta Unariánov, dodnes silná, prežila nezjavenie sa medzigalaktickej vesmírnej flotily 27. septembra 1975.

Prečo by sa skupinová viera mala stať *silnejšou* po stretnutí sa so zdrvujúcou protiindíciou?

Zvyčajné vysvetlenie tohto javu je založené na kognitívnej disonancii. Keď ľudia urobia „nezvratné“ činy v službe danej viery – rozdajú všetok svoj majetok v očakávaní pristátia lietajúcich tanierov – nemôžu si pripustiť, že sa mýlili. Tlak na ich názor vytvára obrovskú kognitívnu disonanciu; musia nájsť posilňujúce myšlienky, aby tento šok prekonali, a tak sa stanú ešte väčšími fanatikmi. Podľa tohto výkladu je silnejší skupinový fanatizmus výsledkom zvýšenia fanatizmu jednotlivcov.

Pozeral som na Java applet, ktorý znázorňoval využitie ochladzovania vyparovaním na tvorbu Boseho-Einsteinovho kondenzátu, keď mi napadlo, že na zvýšenie fanatizmu môže pôsobiť celkom iná sila. Ochladzovanie vyparovaním vytvára bariéru potenciálnej energie okolo skupiny horúcich atómov. Tepelná energia je v prírode v podstate štatistická – nie všetky atómy sa pohybujú rovnakou rýchlosťou. Kinetická energia jednotlivých atómov sa mení ako do seba atómy navzájom narážajú. Ak nastavíte bariéru potenciálnej energie len o trochu vyššie než je priemerná teplotná energia, jednotlivé atómy z času na čas získajú dostatočne vysokú kinetickú energiu na únik z pasce. Keď unikne nezvyčajne rýchly atóm, odnáša so sebou nezvyčajne veľké množstvo kinetickej energie, takže priemerná energia klesne. Skupina sa postupe stáva chladnejšou než je bariéra potenciálnej energie okolo nej. Hranie sa s Java appletom mi to objasnilo.

Vo Festingerovej klasike: „Keď proroctvo zlyhá“ jeden z členov sekty vyšiel z dverí okamžite po tom, čo lietajúci tanier nepristál.[137] Kto sa nahnevá a odíde *prvý*? *Priemerný* člen sekty? Alebo relatívne skeptickejší člen, ktorý mohol dovtedy konať ako hlas umiernenosti, brzda voči fanatickejším členom?

Keď členovia s najväčšou kinetickou energiou uniknú, zvyšné diskusie budú medzi extrémnymi fanatikmi na jednej strane a o trochu menej extrémnymi fanatikmi na druhej strane, pričom konsenzus skupiny bude niekde „uprostred“.

A čo by bola analógia kolapsu, ktorý vytvorí Boseho-Einsteinov kondenzát? Nuž, nemusíme túto analógiu ťahať až tak ďaleko. Ale možno si spomeniete, že som použil analógiu štiepnej jadrovej reakcie na afektívnu špirálu smrti; keď skupina vylúči všetky svoje umiernené hlasy, všetci zostávajúci ľudia sa budú navzájom podporovať a potláčať nesúhlas, čo môže zvýšiť priemerný fanatizmus vo vnútri. (Nemám tu žiadnu termodynamickú analógiu, dokiaľ niekto nevyvinie jadrovú bombu, ktorá vybuchne, keď sa ochladí.)

Keď sa Objektivisti dozvedeli o dlhodobej afére Ayn Rand s Nathanielom Brandenom, významná časť Objektivistov sa odtrhla a pridala sa k Brandenovi vyhlasujúc „otvorený systém“ Objektivizmu, ktorý nebude natoľko pevne spojený s Ayn Rand. Kto zostal s Ayn Rand aj po vypuknutí škandálu? Tí, ktorí v ňu *naozaj*, *naozaj* verili – a možno pár nerozhodných, ktorí po odchode umiernených hlasov

→ http://lesswrong.com/lw/lo/uncritical_supercriticality/

137 Leon Festinger, Henry W. Riecken, and Stanley Schachter, *When Prophecy Fails: A Social and Psychological Study of a Modern Group That Predicted the Destruction of the World* (Harper-Torchbooks, 1956).

počuli argumenty iba z jednej strany. To môže vysvetľovať, preto je Inštitút Ayn Rand (údajne) po rozchode ešte fanatickejší než pôvodná zakladajúca skupina Objektivistov okolo Brandena a Rand.

Pred pár rokmi som bol v transhumanistickej mailovej diskusii, kde malá skupina presadzovala „sociálne demokratický transhumanizmus“ žlčovito urážajúc každého libertariána v diskusii. Väčšina libertariánov opustila mailovú diskusiu a väčšina zostávajúcich prestala písať. V dôsledku toho sa zostávajúca skupina posunula výrazne doľava. Bolo to úmyselné? Pravdepodobne nie, pretože si nemyslím, že páchatelia poznali psychológiu tak dobre. (Keď už sme pri tom, nespomínam si, že by som niekde inde videl analógiu s ochladzujúcim odparovaním, hoci to neznamená, že si to niekto nevšimol predtým.) Nanajvýš sa mohli pokúšať stať „väčšími rybami v malom rybníku“.

To je dôvod, prečo je dôležité byť naklonený v prospech tolerovania nesúhlasu. Počkajte ešte dosť dlho *potom*, čo sa vám zdá oprávnené vylúčiť člena zo skupiny, než ho naozaj vylúčite. Keď sa zbavíte starých extrémov, pozícia skupiny sa posunie a niekto iný sa stane čudákom. Ak vylúčite aj jeho, ste na dobrej ceste, aby ste sa stali Boseho-Einsteinovým kondenzátom a, ehm, vybuchli.

Odvrátená strana: Thomas Kuhn veril, že veda sa musí stať „paradigmou“ so spoločným technickým jazykom vylučujúcim nečlenov, kým niečo naozaj dokáže urobiť. V počiatočných štádiách vedy, podľa Kuhna, sa prívrženci snažia zo všetkých síl, aby ich práca bola zrozumiteľná mimo akadémie. Lenže (podľa Kuhna) veda môže urobiť skutočný pokrok ako technická disciplína iba keď sa vzdá podmienky dostupnosti zvonka, a keď vedci pracujúci v danej paradigme začnú vo svojej komunikácii predpokladať oboznámenosť s množstvom základného technického materiálu. Znie to cynicky v porovnaní s tým, čo sa zvyčajne hovorí o vnímaní vedy verejnosťou, ale rozhodne tu vidím zrnko pravdy.

Moja vlastná teória moderovania internetu znie, že musíte byť ochotní vylúčiť trollov a spam, aby vznikla konverzácia. Musíte byť dokonca ochotní vylúčiť milých alebo technicky neinformovaných ľudí z technickej mailovej diskusie, ak chcete, aby sa urobilo niečo užitočné. Skutočne otvorená diskusia na internete rýchlo zdegeneruje. Podľa tejto teórie by ste sa však mali vyvarovať vylučovania *sčítaných* trollov – plnia skrytú funkciu legitimizovania menej extrémneho nesúhlasu. Nemali by ste však mať toľko sčítaných trollov, že sa začnú hádať jeden s druhým alebo začnú dominovať v diskusii. Ak máte jedného človeka, ktorý je známy Chlapík, Ktorý Nesúhlasí So Všetkým, každý s rozumnejším a umiernenejším nesúhlasom nebude vyzerať ako jediný vyčnievajúci klinec. Táto teória internetového moderovania mi možno v praxi až tak dobre nefungovala, takže ju berte s nadhľadom.

⁂→

110. Keď sa nikto neodváži naliehať na zdržanlivosť

Jedného dňa som ráno vstal z postele, zapol som si počítač a môj mailový klient Netscape mi automaticky stiahol prehľad správ na daný deň. V ten konkrétny deň boli správy o tom, že dve unesené lietadlá narazili do World Trade Center.

Toto boli moje prvé tri myšlienky, v uvedenom poradí:

Tuším naozaj žijem v Budúcnosti.

Ako dobre, že to nebol jadrový útok.

a potom

Prehnaná reakcia na toto bude desaťkrát horšia než pôvodná udalosť.

Samotný činiteľ „desaťkrát horšia“ sa ukázal ako hrubé podcenenie. Ani ja som neuhádol, ako zle sa veci vyvinú. V tom je tá náročnosť pesimizmu: je *naozaj ťažké* mieriť dostatočne nízko, aby ste boli príjemne prekvapení zhruba rovnako často ako nepríjemne prekvapení.

→ http://lesswrong.com/lw/lr/evaporative_cooling_of_group_beliefs/

Napriek tomu som si hneď uvedomil, že každý všade bude hovoriť o tom, aká hrozná, aká príšerná táto udalosť bola; a že sa nikto neopováži byť hlasom zdržanlivosti, primeranej reakcie. Prvé odhady 11. septembra boli, že zomrelo šesť tisíc ľudí. Každý politik, ktorý by povedal: „6000 smrtí je 1/8 každoročných amerických obetí dopravných nehôd“ by musel do hodiny podať demisiu.

Nie, 11. september nebol dobrý deň. Ale ak *každý* dostane hviezdičku za zdôrazňovanie, ako veľmi to bolí, a *nikto* sa neopováži naliehať na zdržanlivosť pri protiútoku, potom bude reakcia väčšia, než je primerané, bez ohľadu na to, aká presne by bola primeraná úroveň.

Toto je ešte temnejší zrkadlový obraz šťastnej špirály smrti – špirála nenávisti. Každý, kto napadne Nepriateľa je vlastenec; ktokoľvek sa snaží analyzovať čo len jediné negatívne tvrdenie o Nepriateľovi, je zradca. Ale tak ako je prevažná väčšina všetkých zložitých výrokov nepravdivá, aj prevažná väčšina negatívnych vecí, ktoré môžete o hocikom povedať, dokonca aj o tom najhoršom človeku na svete, je nepravdivá.

Myslím si, že najlepšou ilustráciou bude: „samovražední únoscovia boli zbabelí“. Trocha zdravého rozumu, prosím? Vyžaduje to trocha odvahy dobrovoľne napáliť svojím lietadlom do budovy. Pri všetkých ich hriechoch, zbabelosť na to zozname nebola. Ale predpokladám, že hocičo zlé, čo poviete o teroristovi, bez ohľadu na to, aké je to hlúpe, musí byť pravda. Dostal by som ešte väčšiu hviezdičku, keby som obvinil Al Kaidu, že zavraždila Kennedyho? Alebo keby som ich obvinil, že sú stalinisti? Vážne, *zbabelosť*?

Áno, záleží na to, že únoscovia z 11. septembra neboli zbabelci. Nie iba kvôli realistickému porozumeniu nepriateľovej psychológie. Špirály nenávisti jednoducho spôsobujú priveľa škody. Je skrátka príliš nebezpečné, aby niekde na svete existoval cieľ, či už sú to židia alebo Adolf Hitler, o ktorom *povedať negatívne veci* je dôležitejšie než *povedať presné veci*.

Keď obranné sily obsahujú tisíce lietadiel a stovky tisícov ťažko ozbrojených vojakov, človek by mal vziať do úvahy, že samotný tento imunitný systém je schopný napáchať omnoho väčšie škody než 19 chlapov a štyri civilné lietadlá. USA minuli miliardy dolárov a tisíce životov vojakov, aby sami sebe uškodili účinnejšie, než by mohla snívať ľubovoľná teroristická skupina.

Keby USA úplne ignorovali útok 11. septembra – iba by pokrčili plecami a znovu postavili tú budovu – bolo by to lepšie než skutočný vývoj histórie. Ale takáto politická možnosť nebola. Aj keby sa každý v súkromí dovtípil, že imunitná reakcia bude škodlivejšia než choroba, americkí politici nemali inú možnosť ako si udržať kariéru, než kráčať priamo do pasce Al Kaidy. Ktokoľvek argumentuje za väčšiu reakciu, je vlastenec. Ktokoľvek analyzuje vlastenecké tvrdenie, je zradca.

Na začiatku boli múdrejšie reakcie na 11. september, než by som odhadoval. Videl som niekoho v kongrese – nepamätám sa, koho – povedať pred kamerami: „Zabudli sme, že prvoradou úlohou vlády nie je ekonomika, ani zdravotná starostlivosť, ale ochrana krajiny pred útokom.“ Až sa mi oči rozšírili, že politik dokázal povedať niečo, čo nebol signál na potlesk. Daná osoba z kongresu musela byť vo veľkom emocionálnom šoku, že povedala niečo také... skutočné.

Ale za dva dni skutočný šok pominul a starosť o imidž získala úplnú kontrolu nad politickou debatou. Potom špirála stupňovania celkom prevládla. Akonáhle sa zdržanlivosť stala nevysloviteľnou, bez ohľadu na to, kde diskusia začala, úroveň zúrivosti a hlúposti mohla časom iba rásť.

*→

111. Pokus v Robbers Cave

Mali ste niekedy ako dieťa podozrenie, či váš nejapný „letný tábor“ v skutočnosti neslúži nejakému prefíkanému skrytému účelu – napríklad, či to celé nie je vedecký pokus a „táboroví vedúci“ nie sú v skutočnosti výskumníci pozorujúci vaše správanie?

Ani ja.

→ http://lesswrong.com/lw/ls/when_none_dare_urge_restraint/

Ale boli by sme paranoidnejší, keby sme čítali Konflikt a spolupráca v skupine: Pokus v Robbers Cave od Sherifa, Harveya, Whitea a Sherifa.[138] V tejto skupine pokusné osoby – prepáčte, „táborníci“ - boli 22 chlapci medzi piatym a šiestym ročníkom, vybraní z 22 rôznych škôl v Oklahoma City, zo stabilných protestantských rodín strednej triedy, s dobrým prospechom, mediánové IQ 112. Boli natoľko dobre vychovaní a navzájom podobní, ako len výskumníci dokázali zabezpečiť.

Tento pokus, urobený po zmätenom závere druhej svetovej vojny, mal za úlohu skúmať príčiny – a možné lieky – konfliktu medzi skupinami. Ako vyvolali konflikt medzi skupinami, ktorý mohli skúmať? Nuž, rozdelili týchto 22 chlapcov na dve skupiny po 11 táborníkov a...

...a ukázalo sa, že to úplne stačí.

Výskumníci pôvodne plánovali, že pokus vykonajú v troch etapách. V Etape 1 by sa každá skupina táborníkov usadila zvlášť, nevediac o sebe navzájom. Na konci Etapy by sa skupiny navzájom postupne dozvedeli o svojej existencii. V Etape 2 by pomocou série súťaží a pretekov vyvolali rozpor medzi oboma skupinami.

Nebola potrebná žiadna Etapa 2. Vzájomná nevraživosť existovala prakticky od okamihu, keď sa každá skupina dozvedela o existencii tej druhej: Oni používajú *naše* táborisko, *naše* baseballové ihrisko. Pri prvom stretnutí si obe skupiny začali navzájom nadávať. Dali si mená Štrkáči a Orli (dokiaľ si mysleli, že sú jedinou skupinou v táborisku, nepotrebovali mať meno).

Keď boli v súlade s vopred určenou experimentálnou procedúrou ohlásené súťaže a preteky, súperenie medzi skupinami dosiahlo horúčkovitú úroveň. Prvé dva dni bolo v súťaži jasné športové správanie, ale rýchlo sa vytratilo.

Orli ukradli Štrkáčom vlajku a spálili ju. Štrkáči sa vlámali do chaty Orlov a ukradli modré rifle vodcu skupiny, nafarbili ich na oranžovo a ďalší deň niesli ako vlajku s nápisom „Posledný Orol“. Orli podnikli odvetné vlámanie k Štrkáčom, prevrátili im postele a rozhádzali špinu. Potom sa vrátili do svojej chaty, kde sa zabarikádovali a pripravili si zbrane (ponožky naplnené kameňmi) pre prípad protiútoku. Keď Orli vyhrali záverečný pretek naplánovanej Etapy 2, Štrkáči sa vlámali do ich chaty a ukradli ceny. Toto prerástlo do pästného súboja, ktorý musel personál zastaviť v obave zo zranení. Orli si tento príbeh medzi sebou rozprávali a premenili celú záležitosť na veľkolepé víťazstvo – hnali Štrkáčov „až za polcestu do ich chaty“ (čo nebola pravda).

Každá skupina si vytvorila negatívny stereotyp o Cudzích a kontrastujúci pozitívny stereotyp o Našich. Štrkáči veľa nadávali. Keď Orli vyhrali jednu hru, usúdili, že vyhrali vďaka svojim modlitbám a že Štrkáči prehrali pretože stále nadávali. Orli sa rozhodli, že sami prestanú nadávať. Ďalej usúdili, že keďže Štrkáči stále nadávajú, bude múdrejšie nerozprávať sa s nimi. Orli si vytvorili sebaobraz správnych a slušných; Štrkáči si vytvorili sebaobraz silných a drsných.

Členovia skupín sa držali za nos, keď okolo nich išli členovia druhej skupiny.

V Etape 3 sa výskumníci pokúšali zredukovať napätie medzi dvoma skupinami.

Púhy kontakt (prítomnosť bez súťaženia) nezmenšoval napätie medzi skupinami. Spoločná účasť na príjemnej udalosti – napríklad strieľanie ohňostrojov na štvrtého júla – nezmenšila napätie; namiesto toho sa začali ohadzovať jedlom.

Čo myslíte, že *zabralo*?

(Priestor pred prezradením pointy...)

Chlapcom povedali, že možno bude v celom tábore nedostatok vody, pretože vodný systém má záhadné ťažkosti – možno kvôli vandalom. (Vonkajší nepriateľ, jeden z najstarších trikov.)

V oblasti medzi táborom a nádržou mali strážiť štyri hliadky. (Zo začiatku sa hliadky skladali iba z členov rovnakej skupiny.) Všetky hliadky sa mali stretnúť pri nádrži, ak nič nenašli. Keďže nič nenašli, skupiny sa stretli pri nádrži a uvideli, že z kohútika netečie voda. Obe skupiny chlapcov diskutovali,

138 Muzafer Sherif et al., „Study of Positive and Negative Intergroup Attitudes Between Experimentally Produced Groups: Robbers Cave Study,“ [Štúdia o pozitívnych a negatívnych vzťahoch medzi experimentálne vytvorenými skupinami: štúdia Robbers Cave] Neuverejnený rukopis (1954).

v čom môže byť problém, klopali na boky nádrže, objavili rebrík na vrch, skontrolovali, že vodná nádrž je plná, a nakoniec našli vrece napchané v kohútiku. Všetci chlapci sa zhromaždili okolo kohútika, aby ho vyčistili. Padali návrhy od členov oboch skupín a chlapci z oboch skupín sa ich snažili realizovať.

Keď bol kohútik konečne vyčistený, Štrkáči, ktorí mali poľné fľaše, nenamietali, aby sa Orli napili z kohútika ako prví (Orli so sebou nemali fľaše). Nepadali žiadne urážky, dokonca ani zvyčajné „dámy majú prednosť“.

To nebol koniec súperenia. Ďalšie ráno bolo ďalšie ohadzovanie jedlom s urážkami. Ale niekoľko ďalších spoločných úloh, vyžadujúcich spoluprácu medzi oboma skupinami – napríklad roztláčanie nákladného auta – to dosiahlo. Na konci tábora Štrkáči použili 5 dolárov získaných v pretekoch hádzania fazulí na nákup cukríkov pre všetkých chlapcov z oboch skupín.

Pokus v Robbers Cave vykresľuje psychológiu báND lovcov-zberačov, opakujúcu sa v čase, najlepšie zo všetkých pokusov, ktoré kedy sociológovia vymysleli.

Akúkoľvek podobnosť s modernou politikou si iba predstavujete.

(Niekedy si myslím, že druhá najdôležitejšia vec, ktorú ľudstvo potrebuje, je superzloduch. Možno sa podujmem na túto úlohu, keď dokončím svoju terajšiu prácu.)

* →

112. Každá kauza chce byť sektou

Cade Metz v *The Register* nedávno tvrdil, že tajná mailová diskusia hlavných administrátorov Wikipédie začala byť posadnutá blokovaním všetkých kritikov a možných kritikov Wikipédie. Vrátane zablokovania produktívneho používateľa, keď jeden administrátor – iba *kvôli* jeho produktivite – získal presvedčenie, že tento používateľ je špiónom poslaným z *Wikipedia Review*. A že ľudia navrchu Wikipédie uzavreli svoje rady, aby sa tak bránili. (Zatiaľ som sa tieto obvinenia osobne neskúmal.)

Existuje nejaká hlboká morálna chyba v snahe systematizovať poznanie sveta, ktorý by viedla prívržencov tejto Kauzy do šialenstva? Možno iba ľudia, ktorí majú vrodené totalitárske sklony, sa pokúšajú stať svetovou autoritou na všetko...

Pozor, sklon prisudzovať! (Sklon prisudzovať: robenie záverov o niekoho jedinečných predpokladoch na základe správania, ktoré sa dá plne vysvetliť situáciou, v ktorej nastalo. Keď vidíme, ako niekto kope do predajného automatu, myslíme si, že je to „zurvalec“, ale keď my kopneme do predajného automatu, je to preto, lebo nám zmeškal autobus, ušiel vlak, a tento stroj nám zožral peniaze.) Ak sú obvinenia Wikipédie pravdivé, dajú sa vysvetliť *obyčajnou* ľudskou povahou, nie *mimoriadnou* ľudskou povahou.

Rozdelenie na našich a cudzích je súčasťou bežnej ľudskej povahy. Rovnako ako šťastné špirály smrti a špirály nenávisti. Vznešená Kauza nevyžaduje skrytú chybu svojich prívržencov, aby vytvorila sektársku skupinu. Stačí, že jej prívrženci sú ľudia. Všetko ostatné nasleduje prirodzene, štandardný úpadok, ako keď sa vám chladničke pokazí jedlo, keď vypadne elektrina.

V tom istom zmysle, ako sa každý teplotný rozdiel chce vyrovnať a každý počítačový program sa chce stať zbierkou ad-hoc záplat, každá Kauza sa *chce* stať sektou. Je to stav vysokej entropie, ku ktorému má systém sklon, atraktor v ľudskej psychológii. Nemusí to mať nič spoločné s tým, či je Kauza naozaj Vznešená. Možno si myslíte, že Dobrá Kauza nalepí svoje dobro na všetky stránky ľudí, ktorí sú s ňou spojení – že nasledovníci tejto Kauzy budú mať menší sklon k súbojom o spoločenské postavenie, k rozlišovaniu našich a cudzích, k afektívnym špirálam, k uctievaniu vodcov. Ale veriť jednej pravdivej myšlienke nevypína efekt svätožiary. Vznešená kauza neurobí zo svojich stúpencov niečo iné ako ľudí. Existuje veľa zlých myšlienok, ktoré môžu narobiť mnoho škody – ale to nemusí byť nutne to, o čo tu ide.

Každá skupina ľudí s nezvyčajným cieľom – dobrým, zlým, či hlúpym – sa bude blížiť sa k sektárskemu atraktoru, pokiaľ nevyvinú sústavné úsilie zabrániť tomu. Môžete udržať svoj byt

→ http://lesswrong.com/lw/lt/the_robbers_cave_experiment/

chladnejší než vonkajšie počasie, ale musíte mať stále zapnutú klimatizáciu, a akonáhle vypnete elektrinu – vzdáte boj proti entropii – veci sa vrátia do „normálneho“ stavu.

V jednom zaujímavom prípade bola skupina, z ktorej sa stala polosekta, ktorej bojový pokrik bol: „Racionalita! Rozum! Objektívna skutočnosť!“ (Viac o tomto neskôr.) Dať svojej Veľkej Myšlienke nápis „rozum“ vám nepomôže o nič viac než nalepiť na svoj dom nápis „Zima!“ Aj tak musíte zapnúť klimatizáciu – vynaložiť požadovanú energiu na jednotku času, aby ste obrátili prirodzený úpadok do sektárstva. Uctievanie rozumu vás neurobí príčetnejšími o nič viac, než vám uctievanie gravitácie dovolí lietať. Nemôžete sa rozprávať s termodynamikou, ani sa modliť k teórii pravdepodobnosti. Môžete ich *použiť*, ale nie stať sa ich členmi.

Sektárstvo je kvantitatívne, nie kvalitatívne. Otázka neznie: „Sekta, áno alebo nie?“ ale „Koľko sektárstva a kde?“ Ešte aj vo Vede, ktorá je archetypálnou Pravou Naozaj Vznešenou Kauzou, môžeme pohotovo ukázať na súčasnú bojovú líniu vojny proti sektárskej entropii, kde sa bojová línia pomaly posúva dopredu a dozadu. Sú odborné časopisy ochotnejšie prijať články so známejším autorským rukopisom alebo od neznámeho autora z dobre známej inštitúcie, v porovnaní s neznámym autorom z neznámej inštitúcie? Nakoľko článku veríme na základe autority a nakoľko na základe experimentu? Ktoré odborné časopisy používajú anonymných recenzentov, a nakoľko efektívne je anonymné recenzovanie?

Citujem tento príklad, namiesto štandardného hmlistého obvinenia: „Vedci nie sú otvorení novým myšlienkam,“ pretože ukazuje *bojovú líniu* – miesto, kde ľudskú psychológiu aktívne vytláčame, kde je zhromaždená sektárska entropia pumpovaná von. (Čo samozrejme vyžaduje nejaké tepelné straty.)

Tento článok nie je katalógom techník na aktívne pumpovanie proti sektárstvu. Niektoré také techniky som už spomenul, iné spomeniem neskôr. *Tu* chcem ukázať na to, že hodnota samotnej Kauzy neznamená, že môžete vynaložiť *menej* úsilia na odolávanie sektárskemu atraktoru. A že ak dokážete ukázať prstom na bojovú líniu, neznamená to, že priznávate nehodnosť svojej Vznešenej Kauzy. Možno si myslíte, že keby otázka znela: „Sekta, áno alebo nie?“, boli by ste povinní odpovedať: „Nie“, inak by ste zradili svoju milovanú Kauzu. Ale to je ako keby ste si mysleli, že máte rozdeliť stroje na „dokonale efektívne“ a „neefektívne“ namiesto merania ich strát.

Naopak, ak veríte, že Nečistá Podstata všetkých Hlúpych Iných Káuz spôsobila, že dopadli zle, ak sa smejete na hlúposti „obetí siekt“, ak si myslíte, že zakladatelia a členovia siekt sú mutanti, potom nebudete vynakladať dostatočné úsilie na to, aby ste pumpovali proti entropii – aby ste odolávali svojej ľudskosti.

*→

113. Strážcovia pravdy

Voči racionalistom je občas namierená táto kritika: „Inkvizítori si mysleli, že *oni* majú pravdu! Je jasné, že táto vec s ‚pravdou‘ je nebezpečná.“

Existuje veľa jasných odpovedí, napríklad: „Ak si myslíš, že poznať pravdu *by* ti dalo právo mučiť a zabíjať, robíš chybu, ktorá nijako nesúvisí s epistemológiou.“ Alebo: „A toto historické tvrdenie, ktoré si práve povedal o inkvizícii – to je pravda?“

Obrátená hlúposť nie je inteligencia: „Ak váš terajší počítač prestane fungovať, nemôžete dôjsť k záveru, že všetko na vašom terajšom systéme je pokazené a že potrebujete nový systém bez procesora AMD, bez videokarty ATI... aj keď váš terajší počítač všetko toto má a nefunguje. Možno len potrebujete nový kábel od napájania.“ Na dôjdenie k zlému záveru stačí jeden chybný krok, nemusí byť každý krok chybný. Inkvizítori verili, že 2 + 2 = 4, ale to nebolo zdrojom ich šialenstva. Možno teda ani epistemologický realizmus nebol ten problém?

→ http://lesswrong.com/lw/lv/every_cause_wants_to_be_a_cult/

Vyzerá vieryhodne, že keby sa inkvizícia skladala z relativistov vyznávajúcich, že nič nie je pravda a na ničom nezáleží, že by mali menej nadšenia pre svoje mučenie. Takisto by mali menej nadšenia, keby boli po lobotómii. Myslím, že je to férové prirovnanie.

Ale predsa... si myslím, že vzťah inkvizície k pravde tam zohrával rolu. Inkvizícia verila, že existuje niečo ako pravda, a že je to dôležité; nuž, podobne aj Richard Feynman. Ale inkvizítori neboli hľadači pravdy. Oni boli *strážcovia* pravdy.

Raz som čítal tvrdenie (neviem nájsť zdroj), že dôležitou zložkou *ducha doby* je, či umiestňuje svoje ideály do svojej budúcnosti alebo minulosti. Takmer všetky kultúry pred osvietením verili v Pád z Milosti – že veci boli kedysi vo vzdialenej minulosti dokonalé, ale potom nastala katastrofa a odvtedy ide všetko pomaly dole kopcom:

> V dobe, keď bol život na Zemi úplný... Milovali jeden druhého a nevedeli, že to je „láska k blížnemu“. Nepodvádzali jeden druhého a pritom nevedeli, že sú „dôveryhodní ľudia“. Boli spoľahliví a nevedeli, že to je „dobrá viera“. Žili slobodne spolu, dávajúc a prijímajúc, a nevedeli, že sú štedrí. Z tohto dôvodu ich činy neboli zaznamenané. Nepísali dejiny.
>
> --Cesta Čuang C', preložil Thomas Merton

Dokonalá doba v minulosti podľa našich najlepších antropologických indícií nikdy neexistovala. Ale kultúra, ktorá vidí život ako neúprosne upadajúci, sa veľmi líši od kultúry, v ktorej možno dosiahnuť bezprecedentné výšky.

(Hovorím „kultúra“ a nie „spoločnosť“, lebo v jednej spoločnosti môže byť viacero subkultúr.)

Mohli by ste povedať, že rozdiel medzi napríklad Richardom Feynmanom a inkvizíciou bol ten, že inkvizícia verila, že *má* pravdu, zatiaľ čo Richard Feynman pravdu *hľadal*. To však nie je celkom obhájiteľné, pretože nepochybne existovali pravdy, o ktorých si Richard Feynman tiež myslel, že ich *má*. Napríklad: „Obloha je modrá“ alebo: „2 + 2 = 4“.

Áno, aj vo vede existujú prakticky isté pravdy. Všeobecnú relativitu môže vyvrátiť nejaká budúca fyzika – nie však spôsobom, ktorý by predpokladal, že Slnko bude obiehať okolo Jupitera; nová teória musí starej teórii ukradnúť jej úspešné predpovede, nie im protirečiť. Ale evolučná teória sa nachádza na vyššej organizačnej úrovni než atómy, a nič, čo objavíme o kvarkoch, nevyvráti darwinizmus, alebo biologickú teóriu buniek, alebo chemickú teóriu atómov, alebo stovky ďalších skvelých objavov, ktorých pravda je potvrdená za hranicou *rozumnej* pochybnosti.

Sú toto „absolútne pravdy“? Nie v zmysle, že by mali pravdepodobnosť doslova 1,0. Ale sú to prípady, kde si veda v zásade myslí, že našla pravdu.

A predsa vedci nemučia ľudí, ktorí spochybňujú chemickú teóriu atómov. Prečo nie? Pretože si nemyslia, že im ich istota dáva právo mučiť druhých? No áno, to je rozdiel *na povrchu*; ale prečo si to tí vedci *nemyslia*?

Pretože chémia netvrdí, že existuje nejaký nadprirodzený trest za pochybovanie o chemickej teórii atómov? Ale opäť poďme hlbšie a opýtajme sa: „Prečo?“ Prečo si chemici *nemyslia*, že pôjdete do pekla, ak pochybujete o teórii atómov?

Pretože odborné časopisy by neuverejnili váš článok, dokiaľ by ste nemali solídne experimentálne pozorovanie pekla? Ale príliš mnohí vedci dokážu vôľou potlačiť svoj skeptický reflex. Prečo nemajú chemici súkromné sekty tvrdiace, že nechemici pôjdu do pekla, najmä keď mnohí z nich sú zároveň kresťanmi?

Otázky ako táto nemajú jednoduchú jednofaktorovú odpoveď. Ale tvrdil by som, že *jeden* z faktorov sa týka zastávania *obranného* postoja voči pravde, verzus *produktívneho* postoja voči pravde.

Keď ste Strážcom Pravdy, nemôžete k Pravde prispieť ničím užitočným *okrem* toho, že ju strážite. Keď sa snažíte získať Nobelovu cenu za chémiu objavom nového benzénu alebo fulerénu, niekto, kto spochybňuje teóriu atómov nie je pre váš svetonázor ani tak hrozbou ako stratou času.

Keď ste Strážcom Pravdy, jediné, čo môžete robiť, je odstrkávať nevyhnutný pokles do entropie mlátením do všetkého, čo sa odchýli od Pravdy. Ak existuje nejaký spôsob, ako pumpovať proti entropii,

generovať nové pravdivé názory pri malej tepelnej strate, táto pumpa môže držať pravdu nažive aj bez tajnej polície. V chémii môžete pokusy zopakovať a vidieť sami – a toto udržiava vzácnu pravdu nažive bez potreby násilia.

A nie je to také strašná hrozba, ak sa raz niekde pomýlime – ak na chvíľu uveríme niečomu nepravdivému – pretože *zajtra* môžeme opäť nájsť stratenú zem.

Ale celý tento trik funguje iba preto, lebo experimentálna metóda je „kritériom dobra“, ktoré nie je púhym „kritériom zhody“. Pretože experimenty dokážu obnoviť pravdu bez potreby autority, dokážu aj *prekonať* autoritu a vytvoriť nové pravdivé názory tam, kde predtým žiadne neboli.

Kde existujú kritériá dobra, ktoré nie sú kritériami zhody, tam môžu existovať *zmeny*, ktoré sú *vylepšenia* namiesto *hrozieb*. Kde existujú *iba* kritériá zhody, kde neexistuje spôsob, ako prekonať autoritu, tam neexistuje ani spôsob, ako vyriešiť spor medzi autoritami. Okrem vyhladenia. Väčšia puška vyhráva.

Nechcem týmto ponúkať veľký všeobjímajúci jednofaktorový pohľad na dejiny. Chcem tým ukázať na hlboký psychologický rozdiel medzi vnímaním svojej veľkej úlohy v živote ako *ochranu*, *stráženie*, *uchovávanie*, verzus *objavovanie*, *tvorenie*, *vylepšovanie*. Ukazuje smer „hore“ na stupnici času do minulosti alebo do budúcnosti? Je to rozdiel, ktorý podfarbuje všetko, zapúšťa úponky všade.

Preto som vždy trval na tom, napríklad, že ak chcete začať hovoriť o „etike UI“, mali by ste hovoriť o tom, ako chcete *vylepšiť* dnešnú situáciu pomocou UI, namiesto iba chránenia rôznych vecí pred pokazením. Akonáhle prijmete kritérií púheho súhlasu, začnete strácať zo zreteľa svoje ideály – prestanete vidieť nesprávne a správne, a začnete vidieť jednoducho „iné“ a „rovnaké“.

Ešte by som tvrdil, že tento základný psychologický rozdiel je jedným z dôvodov, prečo akademické kruhy, ktoré prestanú robiť aktívny pokrok, majú sklon stať sa *nevraživými*. (Prinajmenšom podľa jemných štandardov vedy. Ničenie *povesti* je podľa historických štandardov mierne; väčšina defenzívnych systémov viery šla priamo po krku.) Ak veľké otrasy neprichádzajú dosť často na to, aby pravidelne vyzdvihovali mladých vedcov na základe zásluh namiesto konformity, daný odbor prestane odolávať štandardnej degenerácii do autority. Keď sa nerobí veľa objavov, nie je celý deň čo robiť okrem lovenia kacírov.

Aby vaše duševné zdravie malo čo najväčší prospech z postoja objavovať / tvoriť / vylepšovať, musíte *naozaj robiť pokrok*, nielen dúfať, že príde.

*→

114. Strážcovia genofondu

Ako všetci vzdelaní obyvatelia 21. storočia, asi ste počuli o druhej svetovej vojne. Možno si pamätáte, že Hitler a nacisti mali v pláne uskutočniť romantizovaný proces evolúcie, vyšľachtiť novú nadradenú rasu, nadľudí, silnejších a chytrejších než všetko, čo existovalo dovtedy.

V skutočnosti je toto rozšírený omyl. Hitler veril, že árijskí nadľudia *už v minulosti existovali* – severský stereotyp, blonďavé modrooké šelmy – ale boli *znečistení* miešaním sa s nečistými rasami. Bol to rasový Pád z Milosti.

Vypovedá to niečo o tom, nakoľko je západná civilizácia presiaknutá predstavou *pokroku*, keď človek počúva o nacistickej eugenike a počuje: „Pokúšali sa vyšľachtiť nadčloveka.“ *Vy*, milí čitatelia – keby ste na tom *vy* boli tak biedne, že by ste podporovali násilnú eugeniku – *vy* by ste sa pokúsili stvoriť nadčloveka. Pretože umiestňujete svoje ideály do budúcnosti, nie do minulosti. Pretože ste *tvoriví*. Myšlienka vrátiť sa šľachtením späť k nejakému severskému archetypu spred tisíc rokov by vám ani nenapadlo ako možnosť – čo, iba *Vikingovia*? To je *všetko*? Keby ste na tom boli tak biedne, že by ste boli ochotní zabíjať, tak by ste sa sakra pokúsili dosiahnuť výšku, ktorú nikto pred vami nedosiahol, inak by to celé bolo zbytočné, že? Nuž, to je jeden z dôvodov, prečo nie ste nacisti, drahí čitatelia.

→ http://lesswrong.com/lw/lz/guardians_of_the_truth/

Vypovedá to niečo o tom, ako ťažké je pre relatívne zdravého predstaviť si sám seba v koži relatívne chorého, keď počujeme o nacistoch, a prekrútime ten príbeh tak, že z nich urobíme vadných transhumanistov.

Komunisti boli tí vadní transhumanisti. „Nový sovietsky človek“ a také veci. Nacisti boli v tomto príbehu celkom jednoznačne biokonzervatívci.

*→

115. Strážcovia Ayn Rand

> Skeptikom predstava, že rozum môže viesť k sekte, pripadá absurdná. Príznaky sekty sú o 180 naopak oproti rozumu. Ale ja vám predvediem, ako sa to nielenže môže stať, ale sa aj stalo, a to skupine, ktorú by ste mohli považovať za tú najnepravdepodobnejšiu sektu v dejinách. Je to lekcia, čo sa stane, keď sa pravda stane dôležitejšou než hľadanie pravdy.
>
> --Michael Shermer, Najnepravdepodobnejšia sekta v dejinách[139]

Myslím si, že Michael Shermer vysvetľuje Objektivizmus príliš zložito. Dostanem sa k tomu.

Romány Ayn Rand oslavujú technológiu, kapitalizmus, odpor jednotlivca voči Systému, obmedzenú vládu, súkromné vlastníctvo, sebectvo. Jej vrcholný literárny hrdina John Galt bol <PREZRADENIE ZÁPLETKY> vedec, ktorý vynašiel novú formu lacnej obnoviteľnej energie, ale potom ju odmietol dať svetu, pretože zisk by bol rozkradnutý na podporu skorumpovanej vlády. </PREZRADENIE ZÁPLETKY>

A potom sa – nejako – toto celé zvrtlo na morálny a filozofický „uzavretý systém“ s Ayn Rand v strede. Pojem „uzavretý systém“ nie je moje obvinenie; je to pojem, ktorý používa Inštitút Ayn Rand na opísanie Objektivizmu. Objektivizmus je definovaný dielami Ayn Rand. Teraz, keď je Rand mŕtva, Objektivizmus je uzavretý. Ak v ľubovoľnom ohľade nesúhlasíte s dielami Rand, nemôžete byť Objektivista.

Max Gluckman raz povedal: „Veda je každá disciplína, v ktorej hlupáci tejto generácie dokážu prekročiť hranicu, ktorú dosiahli géniovia predchádzajúcej generácie.“ Veda sa posúva dopredu ničením svojich hrdinov, Newton padol za obeť Einsteinovi. Každý mladý fyzik sníva o tom, že bude novým šampiónom, ktorého sa budúci fyzici budú snažiť zvrhnúť z trónu.

Filozofickým vzorom Ayn Rand bol Aristoteles. No, možno bol Aristoteles nádejným mladým matematickým talentom pred 2350 rokmi, ale od jeho čias matematika urobila badateľný pokrok. Bayesovská teória pravdepodobnosti je kvantitatívna logika, v ktorej je Aristotelova kvalitatívna logika iba špeciálnym prípadom; nie je však žiaden náznak, že by Ayn Rand vedela o bayesovskej teórii pravdepodobnosti, keď písala svoj magnum opus, *Atlas pokrčil plecami*. Rand písala o „rozumnosti“, ale neoboznámila sa s moderným výskumom heuristiky a skreslení. Ako môže hocikto tvrdiť, že je majster racionalista, a pritom nevedieť nič o takýchto základných predmetoch?

„Počkaj chvíľu,“ namieta čitateľ, „to nie je fér! *Atlas pokrčil plecami* bol uverejnený v roku 1957! Vtedy ešte prakticky nikto nevedel o Bayesovi.“ Ba. Nabudúce mi poviete, že Ayn Rand zomrela v roku 1982 a nemala šancu prečítať si knihu *Úsudok v neistote: Heuristiky a skreslenia*, ktorá vyšla v tom istom roku.

Veda nie je fér. To som tým chcel povedať. Ašpirujúci racionalista v roku 2007 začína s obrovskou výhodou oproti ašpirujúcemu racionalistovi z roku 1957. Podľa toho poznáme, že nastal pokrok.

Mne osobne myšlienka dobrovoľného prijatia systému vyslovene zviazaného s názormi jedného človeka, ktorý je *mŕtvy*, pripadá niekde medzi hlúpou a samovražednou. Počítač nemá ani päť rokov a už je zastaralý.

→ http://lesswrong.com/lw/m0/guardians_of_the_gene_pool/

139 Michael Shermer, „The Unlikeliest Cult in History,“ *Skeptic* 2, no. 2 (1993): 74–81, http://www.2think.org/02_2_she.shtml.

Tá živosť, ktorú Rand obdivovala na vede, na obchode, na každej železnici, ktorá nahradila cestu pre konské záprahy, na každom mrakodrape postavenom s *novou* architektúrou – to všetko vychádza z princípu *prekonania dávnych majstrov*. Ako môže existovať veda, ak najmúdrejší vedec, ktorý kedy bude, už žil? Kto by mohol postaviť siluetu New Yorku, ktorú Rand tak obdivovala, keby najvyššia budova, aká bude kedy existovať, už bola postavená?

Napriek tomu Ayn Rand neuznávala nikoho väčšieho, ani v minulosti, ani v prichádzajúcej budúcnosti. Rand, ktorá začala obdivom rozumu a jednotlivca, skončila vyháňaním každého, kto sa jej opovážil protirečiť. Shermer: „[Barbara] Branden si spomenula na večer, keď jeden z priateľov Rand poznamenal, že sa mu páči hudba Richarda Straussa. ‚Keď koncom večera odišiel, Ayn povedala, reagujúc svojím čoraz typickejším spôsobom: „Teraz rozumiem, prečo on a ja nikdy nemôžeme byť spriaznené duše. Rozdiel v našich zmysloch života je príliš veľký.“ Často ani nepočkala, kým priateľ odišiel, než urobila takúto poznámku.‘“

Zdá sa, že Ayn Rand sa časom zmenila.

Rand vyrástla v Rusku, a videla boľševickú revolúciu na vlastné oči. Vo veku 21 rokov dostala víza na návštevu príbuzných v Amerike a nikdy sa nevrátila. Je ľahké nenávidieť autoritárstvo, keď ste obeťou. Je ľahké držať palce slobode jednotlivca, keď vy sami ste utláčaní.

Vyžaduje si omnoho silnejšiu povahu báť sa autority, keď máte moc. Keď sa na vás ľudia obracajú s otázkami, je ťažšie povedať?: „Čo do čerta viem ja o hudbe? Som spisovateľka, nie skladateľka,“ alebo: „Je ťažké vidieť, ako obľuba hudobného diela môže byť nepravdivá.“

Keď ste *vy* tá, ktorá utláča tých, čo sa vás opovážia uraziť, uplatňovanie moci vyzerá akosi omnoho *zdôvodniteľnejšie* než keď ste boli tá utláčaná. Akosi vám sami napadajú všetky možné výborné zdôvodnenia.

Michael Shermer ide do detailov, ako si myslí, že Randovej filozofia skončila úpadkom do sektárstva. Shermer napríklad hovorí (pôsobí to tak), že Objektivizmus zlyhal, pretože Rand si myslela, že istota je možná, zatiaľ čo veda si nikdy nie je istá. V tomto so Shermerom nesúhlasím. Chemická teória atómov je sakra istá. Ale chemici sa nestali sektou.

Ja si vlastne myslím, že Shermer sa stal obeťou sklonu prisudzovať, keď predpokladá, že existuje nejaký konkrétny vzťah medzi Randovej filozofiou a spôsobom, ako z jej nasledovníkov stala sekta. Každá kauza sa chce stať sektou.

Ayn Rand ušla zo Sovietskeho Zväzu, napísala knihu o individualizme, ktorá sa mnohým ľuďom páčila, dostala veľa komplimentov a vytvorila si kliku obdivovateľov. Jej obdivovatelia nachádzali krajšie a krajšie veci, ktoré o nej hovorili (šťastná špirála smrti) a jej sa to príliš páčilo, než aby im povedala, nech sklapnú. Zistila, že má moc utláčať tých, ktorí sa jej nepáčili a neodolala pokušeniu moci.

Ayn Rand a Nathaniel Branden mali tajný mimomanželský pomer. (So súhlasom oboch svojich partnerov, čo v mojich očiach znamená veľa. Ak chcete z tohto urobiť „problém“, museli by ste upresniť, že partneri boli *nešťastní* – a ešte aj tak do toho ostatných nič nie je.) Keď sa ukázalo, že Branden „podvádzal“ Rand ešte s inou ženou, Rand sa rozčertila a exkomunikovala ho. Mnohí Objektivisti sa odtrhli, keď sa správa o afére stala verejnou.

Kto zostal s Rand, namiesto nasledovania Brandena alebo celkového opustenia Objektivizmu? Jej *najsilnejší* podporovatelia. Kto odišiel? Predchádzajúce umiernené hlasy. (Ochladzovanie skupinových názorov vyparovaním.) Potom už mala Rand svoju zvyšnú kliku absolútne v hrsti a nebolo dovolené žiadne spochybňovanie.

Jediná nevšedná vec na celej tejto udalosti je, aké všedné to bolo.

Možno si myslíte, že myšlienkový systém, ktorý chváli „rozum“ a „racionalitu“ a „individualizmus“ tým nejako získa nejakú špeciálnu imunitu...?

Nuž, nestalo sa.

Fungovalo to asi rovnako dobre ako dať nápis „Zima“ na chladničku, ktorá nie je zapnutá.

Aktívna snaha potrebná na odolanie sklzu do entropie tam nebola a prirodzene nasledoval rozpad.

A ak toto nazývate „najnepravdepodobnejšia sekta v dejinách“, akurát urážate skutočnosť.

Nech je to lekcia pre nás všetkých: Oslavovať „rozumnosť“ nič neznamená. Dokonca aj povedať: „Musíš zdôvodniť svoje názory Rozumom, nie iba súhlasiť s Veľkým Vodcom“ akurát spustí automatický programček, ktorý vezme všetko, čo Veľký Vodca povedal a vygeneruje zdôvodnenie, ktoré vaši spolunasledovníci budú považovať za Rozum-né.

Kde sa teda dá nájsť skutočné umenie rozumnosti? Matematickým štúdiom teórie pravdepodobnosti a teórie rozhodovania. Vstrebaním kognitívnych vied ako je evolučná psychológia, alebo heuristík a skreslení. Čítaním historických kníh...

„Študujte vedu, nie iba mňa!“ je asi tá najdôležitejšia rada, ktorú Ayn Rand mohla dať svojim nasledovateľom, ale nedala. Niet takého človeka, ani nikdy nebolo, ktorého plecia by uniesli *celú* váhu skutočnej vedy s mnohými prispievateľmi.

Myslím si, že stojí za zmienku, že literárni hrdinovia Ayn Rand boli architekti a inžinieri; John Galt, jej vrcholná postava, bol fyzik; a predsa samotná Ayn Rand nebola veľká vedkyňa. Pokiaľ viem, nebola zvlášť dobrá v matematike. Nemohla dúfať, že bude súperom svojim vlastným hrdinom. Možno práve tam začala strácať zmysel pre vôľu stále sa zdokonaľovať.

Zatiaľ čo ja, viete, obdivujem drzosť Francisa Bacona, ale ponechávam si schopnosť chvastavo priznať: „Keby som sa mohol vrátiť v čase a nejako dosiahnuť, že by Francis Bacon pochopil problém, na ktorom práve pracujem, oči by mu vyskočili z jamiek ako zátky od šampanského a vybuchli by.“

Obdivujem Newtonove úspechy. Môj postoj k volebnému právu pre ženy mi však bráni prijať Newtona ako vzor morálky. Rovnako ako moja znalosť bayesovskej pravdepodobnosti mi bráni vnímať Newtona ako najvyšší neporaziteľný zdroj matematického poznania. A moja znalosť špeciálnej relativity, hoci je slabá a málo používaná, mi bráni vidieť Newtona ako najvyššiu autoritu vo fyzike.

Newton reálne nemal šancu objaviť žiadnu z myšlienok, ktorými sa nad ním vyvyšujem – *ale pokrok nie je fér! To je pointa!*

Veda má hrdinov, ale nemá bohov. Veľké Mená nie sú naši nadriadení, dokonca ani naši súperi, sú to míľniky, okolo ktorých sme prešli na našej ceste; a tým najdôležitejším míľnikom je hrdina, ktorý ešte len príde.

Byť ďalším míľnikom na ceste ľudstva je to najlepšie, čo možno o niekom povedať; ale toto sa Ayn Rand zdalo príliš slabé. A tak sa z nej stal iba Najvyšší Prorok.

* →

116. Dva koany o sektách

Nováčika racionalistu študujúceho u majstra Ougiho pokarhal priateľ slovami: „Celý čas tráviš počúvaním svojho majstra a rozprávaním o ‚rozumnom‘ tomto a ‚rozumnom‘ tamtom – dostal si sa do sekty!“

Nováčika to hlboko znepokojilo. Slová: „Dostal si sa do sekty!“ mu zneli v ušiach, keď si večer ľahol spať, a dokonca aj vo sne.

Ďalší deň nováčik prišiel k Ougimu, vyrozprával mu, čo sa stalo a povedal: „Majster, stále ma trápi obava, že toto celé je naozaj sekta, a že tvoje učenie je iba dogma.“

Ougi odpovedal: „Ak nájdeš na ceste ležať kladivo a predáš ho, môžeš si zapýtať nízku cenu alebo vysokú cenu. Ak si však to kladivo necháš a používaš ho na zatĺkanie klincov, kto môže pochybovať o jeho hodnote?“

Nováčik odpovedal: „Vidíš, toto je presne jedna z tých vecí, ktoré ma znepokojujú – tvoje tajomné zenové odpovede.“

Ougi povedal: „Dobre teda. Budem hovoriť jasne a predložím dokonale rozumné argumenty, ktoré dokážu, že si sa nedostal do sekty. Ale najprv si musíš nasadiť tento hlúpy klobúk.“

Ougi podal nováčikovi obrovský štyridsaťlitrový kovbojský klobúk.

„Ehm, majster...“ povedal nováčik.

→ http://lesswrong.com/lw/m1/guardians_of_ayn_rand/

„Keď ti všetko vysvetlím," povedal Ougi, „uvidíš, prečo je to potrebné. Alebo môžeš v noci naďalej bdieť a uvažovať, či je toto sekta."

Nováčik si nasadil kovbojský klobúk.

Ougi povedal: „Ako dlho budeš opakovať moje slová a ignorovať ich zmysel? Rozhádzané myšlienky začínajú ako pocit pripútanosti k obľúbeným záverom. Si príliš úzkostlivý ohľadom svojho sebaobrazu racionalistu. Prišiel si ku mne hľadať upokojenie. Keby si bol naozaj zvedavý a nevedel by si, či je to tak alebo onak, napadli by ti nejaké spôsoby, ako vyriešiť svoje pochybnosti. Pretože si si potreboval vyriešiť svoju kognitívnu disonanciu, bol si ochotný nasadiť si hlúpy klobúk. Keby som bol zlým človekom, mohol som ťa nechať zaplatiť sto strieborných. Keď sa budeš sústrediť na otázky ohľadom skutočného sveta, cena alebo bezcennosť tvojho chápania bude čoskoro zrejmá. Si ako šermiar, ktorý sa stále obzerá, aby videl, či sa mu niekto neposmieva..."

„To *stačí*," povedal nováčik.

„Ty si chcel dlhšiu verziu," povedal Ougi.

Tento nováčik neskôr nahradil Ougiho a stal sa známym ako Ni no Tachi. Odvtedy nikdy nedovolil svojim žiakom citovať jeho slová v diskusiách, hovoriac: „Používajte techniky, ale nehovorte o nich."

* * *

Nováčik racionalista prišiel k majstrovi Ougimu a povedal: „Majster, obávam sa, že naše racionalistické dódžó je... no... trochu sektárske."

„To je vážna obava," povedal Ougi.

Nováčik chvíľu čakal, ale Ougi nepovedal nič viac.

Nováčik teda opäť prehovoril: „Myslím tým, mrzí ma to, ale keď musíme nosiť tieto rúcha a kapucne – skrátka to vyzerá, akoby sme boli nejakí slobodomurári alebo také čosi."

„Aha," povedal Ougi, „rúcha a kapucne."

„No *áno*, rúcha a kapucne," povedal nováčik. „Skrátka to vyzerá hrozne nerozumne."

„Zodpoviem všetky tvoje obavy," povedal majster, „ale najprv si musíš nasadiť tento hlúpy klobúk." Ougi vytiahol čarodejnícky klobúk, vyšívaný mesiacmi a hviezdami.

Nováčik si vzal klobúk, pozrel naň a potom frustrovane vybuchol: „*A ako môže toto pomôcť?*"

„Keďže sa tak zaujímaš o interakcie medzi odevom a teóriou pravdepodobnosti," povedal Ougi, „nemalo by ťa prekvapiť, že musíš mať špeciálny klobúk, aby si to pochopil."

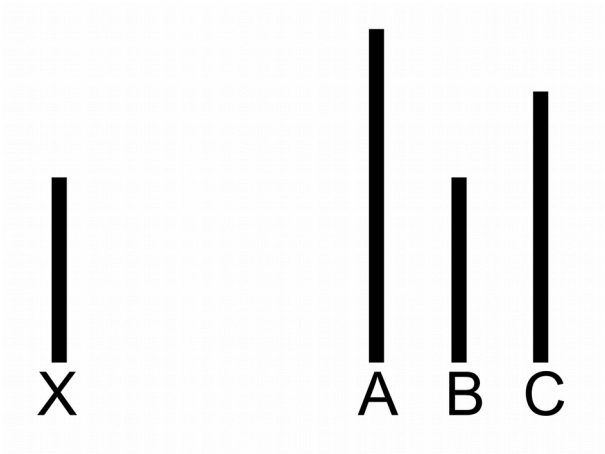
Keď nováčik dosiahol hodnosť doktoranda, prijal meno Bouzo a bol ochotný diskutovať o rozumnosti iba keď nosil šašovský oblek.

*→

117. Aschov pokus o konformite

Solomon Asch, ktorého pokusy začali v 1950-tych rokoch a odvtedy sa mnohokrát replikovali, zvýraznil jav dnes nazývaný „konformita". V klasickom pokuse pokusná osoba vidí úlohu ako na tomto obrázku: Ktorá z čiar A, B, C je rovnako dlhá ako čiara X? Máte chvíľu na to, aby ste určili svoju odpoveď...

→ http://lesswrong.com/lw/m4/two_cult_koans/



Trik je v tom, že pokusná osoba sedí vedľa ďalších ľudí pozerajúcich na obrázok – napohľad ďalšie pokusné osoby, v skutočnosti experimentátorovi spolupracovníci. Ostatné „pokusné osoby“ v tomto pokuse jedna za druhou hovoria, že čiara C vyzerá rovnako dlhá ako čiara X. Skutočná pokusná osoba sedí ako predposledná. Koľko ľudí by v takejto situácii povedalo „C“ - jednoznačne nesprávnu odpoveď, ktorá súhlasí s jednohlasnou odpoveďou ostatných pokusných osôb? Aké percento si myslíte, že by to bolo?

Tri štvrtiny pokusných osôb v Aschovom pokuse dalo „konformnú“ odpoveď aspoň raz. Tretina pokusných osôb odpovedala konformne na viac ako polovicu otázok.

Interview po pokuse ukázalo, že väčšina pokusných osôb tvrdila, že naozaj neveria svojom konformným odpovediam, niektorí však povedali, že si naozaj myslia, že konformná odpoveď bola tá správna.

Ascha tieto výsledky znepokojili:

> Že sme v našej spoločnosti našli taký silný sklon ku konformite... je znepokojujúce. Vyvoláva to otázky o našom spôsobe vzdelávania a o hodnotách, ktorými sa riadi naše správanie.[140]

Nie je to triviálna otázka, či sa pokusné osoby v Aschovom pokuse správali *nerozumne*. Robertova Aumannova veta o súhlase ukazuje, že úprimní Bayesovci sa nemôžu zhodnúť, že sa nezhodnú – ak navzájom vedia o svojich odhadoch pravdepodobnosti, majú rovnaký odhad pravdepodobnosti. Aumannova veta o súhlase bola dokázaná vyše dvadsať rokov po Aschovom pokuse, ale iba formalizuje a posilňuje intuitívne jasnú vec – že názory iných ľudí sú často legitímne indície.

Keby ste sa pozerali na horeuvedený diagram a vedeli by ste *ako fakt*, že zvyšní ľudia v pokuse sú úprimní a vidia ten istý diagram ako vy, a že traja ďalší povedali, že C je rovnako veľké ako X, aká je šanca, že *iba vy* ste jediný, ktorý má pravdu? Ja si nemyslím, že mám v *zrakovom* hodnotení nejakú výhodu – nemyslím si, že by som bol lepší než priemerný človek v posudzovaní, či sú dve čiary rovnako dlhé. Z hľadiska individuálnej rozumnosti dúfam, že by som si všimol svoj veľký zmätok, a potom priradil pravdepodobnosť > 50 % väčšinovému hlasu.

Z hľadiska skupinovej rozumnosti mi pripadá, že úprimný racionalista by mal povedať: „Aké prekvapujúce, že mne *pripadá* B rovnako dlhé ako X. Ale ak sa pozeráme na rovnaký diagram a odpovedáme úprimne, nemám dôvod veriť, že môj odhad je lepší než váš.“ Tá posledná veta je dôležitá – je to omnoho slabšie tvrdenie nesúhlasu než: „Aha, *ja* vidím optický klam – samozrejme rozumiem, prečo si myslíte, že je to C, ale správna odpoveď je B.“

Konformné pokusné osoby v týchto pokusoch nie sú *automaticky* usvedčené z nerozumnosti na základe toho, čo som doteraz opísal. Ale ako ste mohli čakať, čert je skrytý v detailoch výsledkov pokusu. Podľa meta-analýzy vyše sto replikácií od Smitha a Bonda:[141]

140 Solomon E. Asch, „Studies of Independence and Conformity: A Minority of One Against a Unanimous Majority,“ [Štúdie o nezávislosti a konformite: Menšina jedného proti jednohlasnej väčšine] *Psychological Monographs* 70 (1956).

141 Rod Bond and Peter B. Smith, „Culture and Conformity: A Meta-Analysis of Studies Using Asch's (1952b, 1956) Line Judgment Task,“ [Kultúra a konformita: Meta-analýza štúdií požívajúcich Aschovu (1952b, 1956) úlohu hodnotenia čiary] *Psychological Bulletin* 119 (1996): 111–137.

Konformita sa prudko zvyšuje po 3 spolupracovníkoch, ale potom sa ďalej nezvyšuje po 10-15 spolupracovníkoch. Keby boli ľudia rozumne konformní, potom by názor 15 ďalších ľudí mal byť podstatne silnejšou indíciou než názor ďalších 3 ľudí.

Pridanie jediného nesúhlasiaceho – iba jednej osoby, ktorá dá správnu odpoveď, alebo hoci len nesprávnu odpoveď odlišnú od nesprávnej odpovede skupiny – znižuje konformitu *veľmi* prudko, na 5-10 %. Ak používate nejakú intuitívnu verziu Aumannovej vety o zhode a myslíte si, že ak 1 človek nesúhlasí s 3 ľuďmi, tak tí 3 majú pravdepodobne pravdu, potom by ste si vo väčšine prípadov mali myslieť to isté keď 2 ľudia nesúhlasia so 6 ľuďmi. (Nie je to automaticky pravda, ale pravda *ceteris paribus*.) Na druhej strane, ak sú ľudia emocionálne nervózni z toho, že by sami vyčnievali z davu, potom je ľahké ukázať, ako jediný človek, ktorý s vami súhlasí, alebo hoci len jediný človek, ktorý nesúhlasí so skupinou, vás urobí omnoho menej nervóznymi.

Nie je prekvapivé, že pokusné osoby v situácii s jedným nesúhlasiacim si nemysleli, že ich nekonformitu ovplyvnil alebo umožnil tento nesúhlasiaci. Tak ako 90 % šoférov, ktorí si myslia, že patria medzi lepších 50 %, niektorí z nich môžu mať pravdu, ale nie všetci. Ľudia si neuvedomujú príčiny svojej konformity alebo nesúhlasu, čo spochybňuje argument, že sú to prejavy rozumnosti. Napríklad pri hypotéze, že ľudia sa spoločensky-rozumne rozhodujú klamať, aby nevyčnievali z davu, zdá sa, že (aspoň niektoré) pokusné osoby v podmienkach jedného nesúhlasiaceho nepredpokladajú, že by použili takúto „vedomú stratégiu“, keby sa stretli s jednohlasnou opozíciou.

Keď osamelý nesúhlasiaci náhle prepol do *konformity so skupinou*, miera konformity pokusných osôb sa vrátila do rovnakej výšky ako v prípade bez nesúhlasiaceho. Byť prvým nesúhlasiacim je hodnotná (a drahá!) spoločenská služba, ale musíte v nej vydržať.

Konzistentne v rámci pokusov a medzi nimi, skupiny zložené zo samých žien (aj pokusná osoba aj spolupracovníčky) sú omnoho konformnejšie než skupiny zložené zo samých mužov. Asi polovica žien je konformná vo viac než polovici prípadov, oproti tretine mužov. Ak by ste argumentovali, že priemerná pokusná osoba je rozumná, potom sú zrejme ženy príliš prispôsobivé a muži príliš neprispôsobiví, takže žiadna skupina nie je naozaj *rozumná*...

Manipulácie zapadnutia / nezapadnutia do skupiny (napríklad telesne postihnutá pokusná osoba vedľa telesne postihnutých spolupracovníkov) podobne ukázali, že konformita je výrazne silnejšia medzi členmi rovnakej skupiny.

Konformita je nižšia pri očividných diagramoch, ako ten navrchu tejto stránky, oproti diagramom, kde je chyba menej zrejmá. To je ťažké vysvetliť, ak (všetky) pokusné osoby robia spoločensky rozumné rozhodnutie nevyčnievať z davu.

Paul Crowley ma upozornil, že keď pokusné osoby môžu odpovedať spôsobom, ktorý skupina nevidí, konformita tiež klesne, čo je tiež argument proti Aumannovskej interpretácii.

* →

118. O vyjadrovaní svojich obáv

Tá strašidelná vec na Aschovom pokuse s konformitou je, že môžete primäť veľa ľudí, aby povedali, že čierne je biele, ak ich dáte do miestnosti plnej ľudí, ktorí hovoria to isté. Nádejná vec na Aschovom pokuse s konformitou je, že jediný nesúhlasiaci výrazne znížil mieru konformity, dokonca aj keď ten nesúhlasiaci dával inú nesprávnu odpoveď. A *zdrvujúca* vec je, že nesúhlas sa počas pokusu *nenaučil* – keď tento osamelý nesúhlasiaci začal súhlasiť so skupinou, miera konformity sa opäť vrátila.

Byť hlasom nesúhlasu môže skupine priniesť skutočný úžitok. Ale zároveň (je povestné, že) niečo stojí. A potom sa toho musíte držať. Navyše sa môžete mýliť.

Nedávno som mal zaujímavú skúsenosť, keď som začal diskutovať o nejakom projekte s dvoma ľuďmi, ktorí predtým sami robili nejaké plány. Myslel som si, že sú príliš optimistickí a dal som k projektu kopec návrhov typu „pridajme sem pre istotu rezervu“. Čoskoro šiel okolo štvrtý chlap, ktorý

→ http://lesswrong.com/lw/m9/aschs_conformity_experiment/

mal jedného z tých dvoch odviezť domov, a začal dávať návrhy. V tomto bode som mal náhle vhľad, ako sa skupiny stávajú príliš sebaisté, pretože vždy, keď som vzniesol možný problém, ten štvrtý chlap povedal: „Neboj sa, to určite zvládneme!“ alebo niečo podobne ubezpečujúce.

Jednotlivci pracujúci osamote budú mať prirodzene pochybnosti. Budú rozmýšľať: „Dokážem naozaj urobiť XYZ?“, pretože nie je nič neslušné na tom, keď človek spochybňuje *vlastné* schopnosti. Ale keď dvaja neistí ľudia vytvoria skupinu, je slušné hovoriť pekné a ubezpečujúce veci, a neslušné spochybňovať schopnosti toho druhého. Spolu sa stanú optimistickejšími než by bol každý z nich osobitne, pochybnosti každého z nich sa uhasia napohľad dôveryhodným ubezpečením toho druhého, neuvedomujúc si, že ten druhý mal na začiatku rovnaké pochybnosti.

Najdesivejšia možnosť, ktorú naznačuje Aschov pokus s konformitou je prízrak, že každý súhlasí so skupinou, presvedčený sebaistými hlasmi druhých, dávajúc si pozor, aby nedal najavo vlastné pochybnosti – neuvedomujúc si, že ostatní potláčajú podobné obavy. Toto sa nazýva „pluralistická ignorancia“.

Robin Hanson a ja máme dlhodobú diskusiu o tom, kedy presne by sa ašpirujúci racionalisti mali odvážiť nesúhlasiť. Ja sa prikláňam k všeobecne obľúbenému postoju, že nemáte na výber a musíte si tvoriť svoje vlastné názory. Robin Hanson odporúča omnoho rebelskejší názor, že *vy* – nie iba iní ľudia – by ste mali zvážiť, že druhí môžu byť múdrejší. Bez ohľadu na naše rôzne debaty obaja súhlasíme, že Aumannova veta o zhode naznačuje, že spoločné vedomie faktického nesúhlasu ukazuje, že *niekto* musí byť nerozumný. Bez ohľadu na to, aké čudné pohľady na nás vrhajú, stojíme si ohľadom skromnosti za svojím: Zabudnite na to, čo vám ľudia hovoria o individualizme, *mali* by ste venovať pozornosť tomu, čo si druhí ľudia myslia.

Ehm. Pointa je, že pre racionalistu je nesúhlas so skupinou vážna vec. Nemôžete to zmietnuť slovami: „Každý má právo na svoj vlastný názor.“

Myslím si, že najdôležitejšia lekcia, ktorú si môžeme vziať z Aschovho pokusu je rozlišovať medzi „vyjadrením obavy“ a „nesúhlasom“. Upozorniť na vec, ktorú ostatní nevyslovili, neznamená sľúbiť nesúhlas so skupinou na konci debaty.

Ideálny bayesovský proces konvergencie zahŕňa výmenu indícií, ktorá je pre počúvajúceho nepredvídateľná. Aumannova zhoda sa týka iba *spoločných vedomostí*, kde vy viete, ja viem, vy viete, že ja viem, atď. Hansonov článok „Nemôžeme predvídať nesúhlas“ poskytuje obraz, ako zvláštne by vyzeralo sledovať ideálnych racionalistov, ako konvergujú k odhadu pravdepodobnosti; nevyzerá to ako keď sa dvaja obchodníci na trhu dohadujú na cene.

Žiaľ, *spoločensky* príliš nerozlišujeme medzi „vyjadrením obavy“ a „nesúhlasom“. Skupina racionalistov môže súhlasiť, že budú predstierať, že je tam rozdiel, ale ľudia takto nie sú naozaj nastavení. Keď raz prehovoríte, spáchali ste spoločensky neodvolateľný čin; stali ste sa vytŕčajúcim klincom, nesúladom v pohodlnej harmónii skupiny, a to nemôžete odvolať. Každý, koho urazilo vaše vyjadrenie obáv o ich schopnosti úspešne dokončiť úlohu XYZ, bude voči vám dodatočne pravdepodobne chovať rovnaké množstvo trpkosti aj keď nakoniec poviete: „Nie je problém, súhlasím so skupinou.“

Aschov pokus ukazuje, že schopnosť nesúhlasu inšpirovať druhých je skutočná. Aschov pokus ukazuje, že moc konformity je skutočná. Ak sa každý zdrží vyjadrenia svojich súkromných obáv, to skupinu iste povedie do šialenstva. Ale dejiny sú plné príkladov, čo to stojí byť prvým, či dokonca druhým, ktorý povie, že cisár je nahý. Ľudia nemajú ani vrodený sklon rozlišovať medzi „vyjadrením obavy“ a „nesúhlasom so všeobecnou vedomosťou“; tento rozdiel je umelým výtvorom racionalistu. Ak si čítate cynickejší prúd kníh o sebazdokonaľovaní (napríklad Machiavelliho *Vladár*), budú vám radiť celkom maskovať svoju nekonformitu, *nie* vyjadrovať svoje obavy na začiatku a súhlas na konci. Ak preukážete skupine službu, že budete tým, ktorý dá hlas zrejmým problémom, nečakajte, že vám za to skupina poďakuje.

Nesúhlas má svoje ceny a prínosy – či už „nesúhlasíte“ alebo „vyjadrujete obavy“ - a to rozhodnutie je na vás.

*→

119. Osamelý nesúhlas

Aschov pokus s konformitou ukázal, že prítomnosť jediného nesúhlasiaceho dramaticky znížila výskyt „konformných“ nesprávnych odpovedí. Individualizmus je ľahký, ukazuje pokus, keď máte vo svojej odlišnosti spoločnosť. Každá ďalšia osoba v miestnosti, okrem jednej, povie, že čierne je biele. Vy sa stanete druhou osobou, ktorá povie, že čierne je čierne. A cítite sa úžasne: vy dvaja, osamelí a vzdorovití rebeli, proti celému svetu! (Následné interview ukázali, že pokusné osoby v situácii s jedným nesúhlasiacim vyjadrili silný pocit kamarátstva s nesúhlasiacim – hoci si samozrejme nemysleli, že jeho prítomnosť ovplyvnila ich vlastnú nekonformnosť.)

Lenže k rebélii sa môžete *pripojiť* až keď sa niekto niekde stal *prvým* rebelom. Niekto musí povedať, že čierne je čierne po tom, čo počul *všetkých* ostatných, jedného za druhým, povedať, že čierne je biele. A toto – ukazuje pokus – je *omnoho* ťažšie.

Osamelý nesúhlas vám nepripadá ako chodiť do školy oblečený v čiernom. Pripadá vám to ako ísť do školy v šašovskom kostýme.

To je ten rozdiel medzi *pridaním sa k vzbure* a *opustením svorky*.

Ak je nejaká vec, ktorú neznesiem, je to falošnosť – ak nejaký ten čas čítate *Overcoming Bias*, asi ste si to všimli. Nuž, osamelý nesúhlas musí byť jedna z najčastejšie a najostentatívnejšie predstieraných vlastností navôkol. Každý chce byť rebel.

Nechcem tu zhadzovať akt pridania sa k vzbure. Existujú vzbury, ku ktorým sa oplatí pridať. Vyžaduje si odvahu riskovať nesúhlas svojej skupiny rovesníkov, či ešte horšie, ich pokrčenie plecami. Treba však povedať, že ísť na rockový koncert nie je rebelstvo. Ale napríklad vegetariánstvo áno. Ja osobne nie som vegetarián, ale vážim si ľudí, ktorí sú, pretože si to vyžaduje výrazné množstvo tichej odvahy povedať ľuďom, že hamburger na večeru nemôže byť. (Aj keď v okolí San Francisca sa ľudia už bežne opýtajú.)

Ale aj tak, ak poviete ľuďom, že ste vegetarián, budú si myslieť, že rozumejú vašim motívom (aj keď nerozumejú) Môžu nesúhlasiť. Môžu sa uraziť, ak sa vám to podarí oznámiť príliš hrdo, alebo sa môžu uraziť skrátka preto, lebo sú veľmi urážliví. Ale vedia, ako sa k vám majú správať.

Keď niekto chodí do školy v čiernom, učitelia a ostatné deti rozumejú, akú rolu tým chce zaujať v ich spoločnosti. Je to Mimo Systému – veľmi štandardným spôsobom, ktorý každý pozná a chápe. Teda nie, viete, *naozaj* mimo systému. Je to Výzva Štandardnému Mysleniu štandardného typu, takže ľudia pohoršene povedia: „Nerozumiem prečo...“ ale nemusia naozaj myslieť na nič, na čo už nemysleli predtým. Ako sa hovorí: „Už niekedy nejaká ‚podvratná literatúra‘, ktorú ste čítali, spôsobila, že ste zmenili nejaký politický názor?“

Naozajstnú odvahu si vyžaduje priniesť medzi ľudí okolo seba úplné *nepochopenie*, keď urobíte niečo, čo *nie je* iba Štandardná Vzbura číslo 37, niečo, na čo nemajú pripravený scenár. Nehnevajú sa na vás ako na rebela, iba si myslia, že ste akýsi divný a odvrátia sa. Táto predstava vytvára omnoho hlbší strach. To je rozdiel medzi vysvetľovaním vegetariánstva a vysvetľovaním kryoniky. Na svete existujú aj iní kryonici, niekde, ale nie sú tam vedľa vás. Musíte to vysvetliť sám, ľuďom, ktorí si myslia, že ste skrátka *divný*. Nie je to zakázané, ale je to mimo hraníc, o ktorých ľudia ani nerozmýšľajú. Vy si idete nechať zmraziť hlavu? Vy si myslíte, že vás to zachráni pred smrťou? Čo tým myslíte, že informácia v mozgu? He? Čo? *Šibe* vám?

Som v pokušení napísať post facto vysvetlenie pomocou evolučnej psychológie: Mohli ste dať dokopy skupinku kamarátov a odísť od svojej bandy lovcov-zberačov, ale musieť ísť do lesa *sám* bol pravdepodobne rozsudok smrti – prinajmenšom z hľadiska rozmnožovania. Nemyslíme na to explicitne, ale to nie je v povahe evolučnej psychológie. Pridať sa k vzbure, o ktorej každý vie, je strašidelné,

→ http://lesswrong.com/lw/ma/on_expressing_your_concerns/

ale zďaleka nie také strašidelné ako robiť niečo naozaj odlišne. Niečo, čo v pradávnych časoch mohlo skončiť nie rozdelením skupiny, ale vyhnaním vás samého.

Ako dokladá prípad kryoniky, strach myslieť *naozaj* odlišne je silnejší než strach zo smrti. Lovci-zberači museli byť pripravení čeliť smrti každý deň pri love veľkých cicavcov alebo len pri prechádzaní sa vo svete plnom dravcov. Potrebovali tú odvahu aby mohli žiť. Odvaha vzdorovať kmeňovým štandardným spôsobom myslenia, myslieť si veci, ktoré vyzerali naozaj divne – nuž, toto pravdepodobne svojim nositeľom rovnako neprospievalo. Nerozmýšľame nad tým explicitne; tak evolučná psychológia nefunguje. My ľudia sme skrátka postavení takým spôsobom, že viacerí z nás idú skákať padákom než sa prihlásiť na kryoniku.

A to ešte nie je tá najväčšia odvaha. Na svete existuje viac ako jeden kryonicista. Iba Robert Ettinger to musel povedať ako *prvý*.

Aby ste boli revolučným *vedcom*, musíte byť prvým človekom, ktorý bude protirečiť tomu, čo si myslí každý, koho poznáte. To nie jediná cesta k vedeckej sláve; je to pomerne vzácne aj medzi tými veľkými. Nikto sa nemôže stať vedeckým revolucionárom tým, že sa bude snažiť napodobňovať revolucionárov. Môžete sa tam dostať iba nasledovaním správnych odpovedí vo všetkých veciach, či je správna odpoveď revolučná alebo nie. Ale ak postupom času – ak po vstrebaní všetkej moci a múdrosti z vedomostí, ktoré už boli zhromaždené – ak po tom všetkom a s dávkou čistého šťastia zistíte, že vás vaše hľadanie čistej správnosti vedie na nové územie... *vtedy* máte príležitosť prejaviť odvahu zlyhať.

Toto je skutočná odvaha osamelého nesúhlasu, ktorú sa každá poondená rocková kapela na svete pokúša predstierať.

Samozrejme nie všetko, čo si vyžaduje odvahu, je dobrý nápad. Chce to odvahu skočiť z útesu, ale potom sa iba rozpleštíte.

Strach z osamelého nesúhlasu je brzdou k dobrým myšlienkam, ale nie každá nesúhlasná myšlienka je dobrá. Pozrite si od Robina Hansona Proti voľnomyšlienkárom. Väčšina problému ako mať nové pravdivé vedecké myšlienky sa týka časti „pravdivé“.

Naozaj nie je *potrebné* byť odlišný kvôli samotnej odlišnosti. Ak robíte veci odlišne iba keď vidíte mimoriadne dobrý dôvod, stále budete mať viac než dostatok problémov na celý zvyšok života.

Existuje okolo nás niekoľko naozaj nefalšovaných skupín obrazoborcov. Cirkev SubGénia má napríklad asi ako skutočný cieľ *zmiasť* bežných ľudí, nie iba uraziť ich. Existujú aj ostrovy skutočnej tolerancie vo svete, napríklad stretnutia science fiction. *Existujú* istí ľudia, ktorí sa neboja oddeliť od svorky. Takých ľudí je v skutočnosti mnoho menej než tých, čo sa považujú za rebelov; ale existujú. A predsa sú vedeckí revolucionári omnoho zriedkavejší. Rozmýšľajte o tom.

No a *ja*, viete, ja *naozaj som* obrazoborec. Každý si to o sebe myslí, ale ja mám *pravdu*, viete. Ja by som *určite* nosil šašovský kostým do školy. Moje vážne rozhovory boli s knihami, nie s inými deťmi.

Ale ak si myslíte, že by ste *určite* nosili ten šašovský kostým, nebuďte príliš hrdí ani na to! Znamená to iba, že musíte vynaložiť úsilie *opačným smerom*, aby ste sa vyhli príliš ľahkému nesúhlasu. To je to, čo musím robiť ja, aby som napravil svoju povahu. Iní ľudia majú dôvody, prečo si myslia to, čo si myslia, a celkom ich ignorovať je rovnako zlé ako báť sa im protirečiť. Nechceli by ste skončiť ako voľnomyšlienkár. To totiž nie je *cnosť* – iba skreslenie iným smerom.

*→

120. Sektárske antisektárstvo

Vstúpiť do sekty je pravdepodobne jedna z najhorších vecí, ktoré sa vám môžu stať v modernom svete. V najlepšom prípade skončíte v skupine úprimných ale pomýlených ľudí, ktorí robia úprimnú chybu, ale inak sa správajú dobre, miniete kopec času a peňazí a nakoniec vám nezostane nič. Toto je vlastne opis neúspešnej začínajúcej firmy v Silicon Valley. Čo by vraj mala byť čertovsky otrasná skúsenosť, keď sa nad tým zamyslím. Takže áno, veľmi strašidelné.

→ http://lesswrong.com/lw/mb/lonely_dissent/

Skutočné sekty sú nepomerne horšie. „Bombardovanie láskou“ ako náborová technika, zameraná na ľudí, ktorí prekonávajú osobnú krízu. Spánková deprivácia. Vyčerpanosť vyvolaná ťažkou prácou. Ďaleké cestovanie, aby sa členovia izolovali od priateľov a rodiny. Denné stretnutia, kde sa spovedajú nečisté myšlienky. Nie je nezvyčajné, že sekta oberie člena o *všetky* peniaze – životné úspory aj mesačnú výplatu – čím ich núti do závislosti v jedle aj oblečení. Hladovka ako test za neposlušnosť. Vážne vymývanie mozgu a vážne ubližovanie.

Keď zvážim toto všetko, mal by som asi viac súcitiť s ľuďmi, ktorí sú hrozne nervózni z pridania sa k nejakému čudne vyzerajúcemu správaniu, že *možno vstupujú do sekty*. Nemalo by mi to liezť na nervy. Lenže lezie.

Bod číslo jedna: „Sekty“ a „nie sekty“ nie sú prirodzene oddelené druhy ako psy a mačky. Ak sa pozriete na hociktorý zoznam charakteristík sekty, nájdete tam prvky, ktoré by mohli ľahko opisovať politické strany a firmy - „členovia skupiny sú povzbudzovaní, aby nedôverovali kritike zvonku, lebo sú za ňou skryté motívy“, „hierarchické autoritatívna štruktúra“. Už som písal o skupinových zlyhaniach ako je polarizácia skupiny, šťastné špirály smrti, nekritickosť, a ochladzovanie vyparovaním, pričom každé z toho asi prispieva ku všetkým ostatným. Keď sa tieto zlyhania pritočia k sebe a spoja, ich spojením vznikne Super-Zlyhanie hlúpejšie než každá z jeho častí, ako Voltron. Ale to nie je sektárska *podstata*; to je sektársky *atraktor*.

Psy sa rodia so psou DNA a mačky sa rodia s mačacou DNA. V dnešnom svete nie je nič medzitým. (Ani s genetickými manipuláciami by nebolo také ľahké vytvoriť organizmus s polovicou psích génov a polovicou mačacích génov.) Neexistuje žiadna navzájom sa povzbudzujúca množina psích vlastností, ku ktorým sa jedna mačka môže napoly zatúlať a stať sa polopsom.

Ľudská myseľ, ktorá rozmýšľa v kategóriách, asi dáva prednosť podstatám pred atraktormi. Človek chce povedať: „Toto je sekta“ alebo „Toto nie je sekta“ a tým je úloha klasifikácie navždy uzavretá. Ak pozorujete, že Sokrates má desať prstov, nosí šaty a hovorí plynulo po grécky, môžete povedať: „Sokrates je človek“ a z toho odvodiť, že „Sokrates je zraniteľný bolehlavom“ bez toho, že by ste robili konkrétne krvné testy na potvrdenie jeho smrteľnosti. Rozhodli ste Sokratovu ľudskosť raz a navždy.

Ale ak pozorujete, že istá skupina ľudí asi vykazuje polarizáciu na našich a cudzích, a vidí kladný efekt svätožiary okolo svojej Úplne Najobľúbenejšej Veci – čo môže byť Objektivizmus alebo vegetariánstvo alebo neurónové siete – nemôžete *na základe zatiaľ zhromaždenej indície* odvodiť, či dosiahli nekritickosť. Nemôžete odvodiť, či je ich hlavná myšlienka pravdivá, nepravdivá, alebo naozaj užitočná no nie až taká užitočná, ako si myslia. *Z doteraz zhromaždených informácií* nemôžete odvodiť, či sú inak slušní alebo či vás odlákajú do izolácie a budú vám odopierať spánok a jedlo. Vlastnosti sekty nie sú všetky prítomné alebo všetky neprítomné.

Ak si pozriete online hádky „X je sekta“, „X nie je sekta“, jedna strana prechádza cez internetový zoznam vlastností sekty, nájde jednu, ktorá sedí a povie: „Preto je to sekta!“ A obranca nájde vlastnosť, ktorá nesedí a povie: „Preto to nie je sekta!“

Nedokážete zostaviť presný obraz dynamiky uvažovania skupiny pomocou takéhoto esencializmu. Musíte venovať pozornosť jednotlivým vlastnostiam jednotlivo.

Navyše, obrátená hlúposť nie je inteligencia. Ak vás zaujíma ústredná *myšlienka*, nie iba skupina, ktorá ju realizuje, potom múdre myšlienky môžu mať hlúpych nasledovníkov. Mnoho priaznivcov new age rozpráva o „kvantovej fyzike“, ale to nie je dôkaz proti kvantovej fyzike. Samozrejme aj hlúpe myšlienky majú hlúpych nasledovníkov. S binárnym esencializmom sa spája predstava, že ak odvodíte, že skupina je „sekta“, potom jej názory musia byť nepravdivé, pretože nepravdivé názory sú vlastnosťou siekt, tak ako mačky majú srsť. Ak vás zaujíma myšlienka, pozerajte na myšlienku, nie na ľudí. Sektárstvo je skôr vlastnosť *skupín* než *hypotéz*.

Druhá chyba je, že keď sa ľudia nervózne pýtajú: „Toto nie je sekta, však?“, znie mi to akoby hľadali *ubezpečenie o rozumnosti*. Myšlienka, že racionalista by sa nemal príliš pripútať ani k svojmu sebaobrazu racionalistu si zaslúži svoj vlastný článok (ale pozrite si Dvanásť cností, Načo pravda? A..., a Dva koany o sektách). Ale aj bez zachádzania do detailov môže byť človeku jasné, že *nervózne hľadanie ubezpečenia* nie je najlepší stav mysle na vyhodnocovanie otázok o rozumnosti. Nebudete

skutočne zvedaví ani myslieť na to, ako naplniť svoje pochybnosti. Namiesto toho si nájdete nejaký internetový zdroj, ktorý povie, že sekty používajú spánkovú depriváciu na ovládanie ľudí, vy si všimnete, že Vaša Obľúbená Skupina nepoužíva spánkovú depriváciu, a dôjdete k záveru: „Nie je to sekta. Uf!“ Keď to nemá srsť, nemôže to byť mačka. Veľmi ubezpečujúce.

Lenže každá kauza sa chce stať sektou, či už je samotná kauza múdra alebo hlúpa. Rozdeľovanie na našich a cudzích atď. je súčasť ľudskej povahy, nie špeciálna kliatba mutantov. Rozumnosť je výnimka, nie pravidlo. Musíte vynakladať sústavné úsilie na udržanie rozumnosti proti prirodzenému sklzu do entropie. Ak sa rozhodnete: „Nie je to sekta!“ a s úľavou si vydýchnete, potom nebudete vynakladať priebežné úsilie potláčať *prirodzené* sklony k sektárstvu. Rozhodnete sa, že sektárska podstata je neprítomná a prestanete pumpovať proti entropii atraktoru sekty.

Ak ste hrozne nervózni zo siekt, potom budete chcieť popierať každý náznak hocijakej vlastnosti, ktorá pripomína sektu. Lenže *každá* skupina, ktorá má cieľ a vidí ho v pozitívnom svetle, je ohrozená efektom svätožiary a bude musieť pumpovať proti entropii, aby sa vyhla afektívnej špirále smrti. Toto platí aj pre bežné organizácie ako sú politické strany – ľudia, ktorí si myslia, že „liberálne hodnoty“ alebo „konzervatívne hodnoty“ dokážu vyliečiť rakovinu atď. Platí to pre začínajúce firmy v Silicon Valley, neúspešné aj úspešné. Platí to pre používateľov Macu aj pre používateľov Linuxu. Efekt svätožiary nezačne byť okej len preto, že to robí každý; keby každý skočil z útesu, vy by ste neskočili. Proti chybám v uvažovaní treba bojovať, nie tolerovať ich. Ale ak ste príliš nervózni ohľadom: „Ste si *istí*, že toto nie je sekta?“ potom sa budete zdráhať vidieť *akékoľvek* znaky sektárstva, pretože to by naznačovalo, že ste v sekte a *Toto nie je sekta!!!* Nebudete teda vidieť aktuálne bojiská, kde sa bežné sklony k sektárstvu plížia vpred alebo sú zatláčané naspäť.

Tretia chyba nervózneho pýtania sa: „Toto nie je sekta, však?“ je to, že *nervozita* je tam, podľa môjho silného podozrenia, z celkom nesprávnych dôvodov.

Čím to je, že skupiny, ktoré vychvaľujú svoju Šťastnú Vec až ku hviezdam, povzbudzujú členov, aby darovali všetky svoje peniaze a pracovali v dobrovoľnom nevoľníctve, a majú súkromné pozemky, kde sú ich členovia prísne izolovaní, nazývame „náboženstvá“ a nie „sekty“, ak tu už boli niekoľko storočí?

Prečo väčšina ľudí, ktorí sa na kryoniku nervózne pýtajú: „Toto nie je sekta, však?“ by nebola rovnako nervózna z účasti na politickej mítingu republikánov alebo demokratov? Rozdelenie na našich a cudzích a šťastné špirály smrti sa môžu vyskytnúť v politickej diskusii, v bežnom náboženstve, v športovom fandení. Keby *nervozita* pochádzala zo strachu z *chýb rozumu*, ľudia by sa pýtali: „Toto nie je rozdelenie na našich a cudzích, však?“ na politických mítingoch demokratov alebo republikánov rovnako bojazlivým hlasom.

Existujú legitímne dôvody báť sa libertariánov menej než kultu lietajúcich tanierov, pretože libertariáni nemajú povesť, že by používali spánkovú depriváciu na konverziu ľudí. Lenže ani kryonici nemajú povesť používania spánkovej deprivácie. Prečo by ste sa teda mali viac báť toho, že vám zmrazia hlavu, keď prestanete dýchať?

Mám podozrenie, že táto *nervozita* nie je strach z nesprávneho názoru, ani strach z fyzickej ujmy. Je to strach z osamelého nesúhlasu. Ten nervózny pocit, ktorí mali pokusné osoby v Aschovom pokuse s konformitou, keď všetky ostatné pokusné osoby (v skutočnosti spolupracovníci) jedna za druhou hovorili, že čiara C je rovnako dlhá ako čiara X, a pokusnej osobe sa zdalo, že čiara B je rovnako dlhá ako čiara X. Strach z opustenia svorky.

To je dôvod, prečo skupiny, ktorých názory tu boli dostatočne dlho, aby vyzerali „normálne“ nevyvolávajú rovnakú nervozitu ako „sekty“, hoci oficiálne náboženstvá vás môžu tiež obrať o všetky peniaze a poslať vás do kláštora. To je dôvod, prečo skupiny ako politické strany, ktoré sú silne zodpovedné za chyby v rozumnosti, nevyvolávajú takú nervozitu ako „sekty“. Slovo „sekta“ sa nepoužíva, aby symbolizovalo chyby v rozumnosti, používa sa ako nálepka na niečo, čo *vyzerá čudne*.

Nie každá zmena je zlepšenie, ale každé zlepšenie je nevyhnutne zmena. Ak chcete niečo robiť lepšie, nemáte inú možnosť ako robiť to inak. Všeobecná múdrosť obsahuje veľké množstvo skutočnej múdrosti; áno, má zmysel vyžadovať pre podivnosť bremeno dôkazu navyše. Ale táto *nervozita* nie je

nejaké úmyselné, rozumné zvažovanie. Je to strach veriť niečomu, čo spôsobí, že na vás budú vaši priatelia pozerať naozaj čudne. A tak sa ľudia pýtajú: „Toto nie je *sekta*, však?“ tónom, ktorý by nikdy nepoužili pri návšteve politického mítingu ani pri postavení gigantickej vianočnej výstavy.

To je tá časť, ktorá má štve.

Je to akoby, akonáhle veríte niečomu, čomu neverili vaši predkovia, priletela z oblohy Sektárska Víla a naplnila vás Sektárskou Esenciou, a ani sa nenazdáte a už všetci nosíte rúcha a spievate. Akoby „čudné“ názory boli *priamou príčinou* problémov, a nie spánková deprivácia a bitky. Škoda spôsobená sektami – samovražda Heaven's Gate a tak ďalej – iba ukazuje, že každý s čudnými názormi je blázon; prvá a hlavná vlastnosť „člena sekty“ je, že je to Outsider s Podivnými Spôsobmi.

Áno, spoločensky nezvyčajný názor vystavuje skupinu nebezpečenstvu kvôli mysleniu na našich a cudzích a ochladzovaniu vyparovaním a ďalším problémom. Ale nezvyčajnosť je rizikový faktor, nie samotná choroba. To isté platí pre cieľ, o ktorom si myslíte, že stojí za dosiahnutie. Či je tento názor pravdivý alebo nie, mať pekný cieľ vás vždy vystavuje riziku šťastnej špirály smrti. Ale to robí úžasné ciele rizikovým faktorom, nie chorobou. Niektoré ciele naozaj stoja za nasledovanie.

Na druhej strane, nevidím legitímny dôvod na spánkovú depriváciu alebo vyhrážanie sa nesúhlasiacim bitkou, bodka. Keď skupina robí toto, či už ju voláte „sekta“ alebo „nie sekta“, už ste priamo odpovedali na pragmatickú otázku ohľadom toho, či do nej vstúpiť.

Problém číslo štyri: Strach z osamelého nesúhlasu je niečo, čo *samotné sekty* zneužívajú. Strach z toho, že sa na vás vaši priatelia budú pozerať odsudzujúco je *presne ten efekt, ktorý skutočné sekty používajú na konvertovanie a udržanie si členov* – obklopia vás jednohlasnou zhodou medzi veriacimi sekty.

Strach zo zvláštnych myšlienok, impulz ku konformite, nepochybne varoval mnoho potenciálnych obetí pred kultmi lietajúcich tanierov. Keď ste mimo, drží vás mimo. Ale keď ste *vnútri*, drží vás *vnútri*. Konformita vás iba prilepí tam, kde práve ste, či je to dobré alebo zlé miesto.

Človek by chcel, aby existoval nejaký spôsob, ako vedieť *naisto*, že nie ste v „sekte“. Nejaká jednoznačná, zdrvujúca replika pre ľudí, ktorí sa naňho čudne pozerajú. Nejaký spôsob, ako raz a navždy vedieť, že robíte správnu vec, bez tých ustavičných pochybností. Myslím, že sa to volá „potreba uzavretia“. A – samozrejme – aj toto sekty zneužívajú.

Preto to volám „sektárske antisektárstvo“.

Žiť s pochybnosťami nie je cnosť – cieľom každej pochybnosti je zničiť samu seba úspechom alebo zlyhaním, a pochybnosť, ktorá len tak čaká, nedosiahne nič. Ale niekedy pochybnosti chvíľu trvá, než sa zničí. Žiť so zásobníkom momentálne nevyriešených pochybností je pre racionalistu nevyhnutný životný fakt. Pochybností by sme sa nemali báť. Inak si budete musieť vybrať medzi veľmi strašidelným a veľmi hlúpym životom.

Ak naozaj, úprimne neviete zistiť, či daná skupina je „sekta“, potom si musíte vybrať v podmienkach neistoty. O tom je celá teória rozhodovania.

Problém číslo päť: Nedostatok strategického myslenia

Poznám ľudí, ktorí si dávajú pozor pri singularitárstve a sú opatrní *aj* pri politických stranách a tradičných náboženstvách. Sú *opatrní*, nie nervózni ani defenzívni. Títo ľudia na prvý pohľad vidia, že singularitárstvo očividne nie je plnohodnotná sekta so spánkovou depriváciou atď. Ale boja sa, že sa singularitárstvo sektou *stane*, kvôli rizikovým faktorom ako premena predstavy mocnej UI na Super Šťastného Činiteľa (činiteľa definovaného v prvom rade tým, že súhlasíte s každou peknou vecou, ktorá sa o ňom povie). Že niečo dnes nie je sekta, ešte neznamená, že sa to nestane sektou v budúcnosti. Sektárstvo je atraktor, nie podstata.

Či ma otravuje *tento* druh opatrnosti? Vôbec nie. Ja sám trávim dosť času znepokojovaním sa nad týmto scenárom. Pokúšam sa svoje kamienky go umiestňovať tak, aby som vopred blokoval pohyb týmto smerom. Preto som napríklad napísal sériu článkov o sektárskych zlyhaniach rozumu.

Ľudia, ktorí hovoria o „rozumnosti“ tiež majú rizikový faktor navyše. Dávať ľuďom rady, ako majú rozmýšľať, je svojou podstatou nebezpečná vec. Ale je to *rizikový faktor*, nie *choroba*.

Obe moje obľúbené Kauzy majú riziko sektárstva. Napriek tomu sa ma pýtajú: „Si si istý, že toto nie je sekta?“ omnoho častejšie, keď hovorím o mocných UI, než keď hovorím o teórii pravdepodobnosti a kognitívnej vede. Neviem, či je jeden rizikový faktor vyšší než druhý, ale viem, ktoré z toho *znie čudnejšie*.

Problém číslo 6 otázky: „Toto nie je sekta, však?“...

Už len tá otázka samotná ma stavia do veľmi otravnej situácie typu Hlava 22. Skutočný Zlý Guru by samozrejme použil nervozitu dotyčného proti nemu samotnému a vytvoril by dôveryhodný prepracovaný argument vysvetľujúci Prečo Toto Nie Je Sekta, a dotyčný by ho ochotne prijal. Občas mám dojem, že presne toto od mňa ľudia *žiadajú*! Kedykoľvek sa pokúšam písať o sektárstve a ako sa mu vyhnúť, mám pocit, akoby som sa poddával tejto pomýlenej túžbe – akoby som v konečnom dôsledku dával ľuďom *ubezpečenie*. Ešte aj keď hovorím ľuďom, že je nevyhnutný ustavičný boj proti entropii.

Cítim sa, akoby som sa robil prvým nesúhlasiacim v Aschovom teste konformity, hovoriac ľuďom: „Áno, čiara X naozaj je rovnaká ako B, je okej ak to poviete aj vy.“ Oni by sa nemali potrebovať pýtať! Alebo ešte horšie, cítim sa, akoby som predkladal prepracovaný argument Prečo Toto Nie Je Sekta. Je to *nesprávna otázka*.

Však sa sami pozrite na myšlienkové procesy skupiny a rozhodnite sa sami, či je to niečo, na čom sa chcete zúčastniť, keď sa konečne zbavíte svojho strachu z nezvyčajnosti. Je to vaša vlastná zodpovednosť prestať myslieť sektársky, bez ohľadu na to, v ktorej skupine momentálne fungujete.

Keď sa niekto opýta: „Toto nie je sekta, však?“, potom bez ohľadu na to, ako odpoviem, sa vždy cítim, akoby som niečo obhajoval. Ja nemám rád tento pocit. Nie je úlohou Bayesovského Majstra dávať ubezpečenia, ani nie je úlohou racionalistu niečo brániť.

Sekty žijú so skupinového myslenia, nervozity, túžby po ubezpečení. Nemôžete nervozitu poslať preč púhym želaním, a falošná sebadôvera je ešte horšia. Ale dokiaľ niekto potrebuje ubezpečenie – hoci len ubezpečenie, že je racionalistom – vždy to bude diera v jeho brnení. Šikovný šermiar sa sústredí na cieľ a neobzerá sa, či sa mu niekto náhodou nevysmieva. Keď viete, čo sa pokúšate urobiť a prečo, viete, či to robíme alebo nie, a či vám skupina pomáha alebo vás brzdí.

(PS: Ak niekto príde za vami a opýta sa: „Si si *istý*, že toto nie je sekta?“, nepokúšajte sa mu všetky tieto pojmy vysvetliť jednou vetou. Podceňujete inferenčné vzdialenosti. Dotyčný povie: „Aha, takže *pripúšťaš*, že ste sekta!“ alebo „Počkaj, ty hovoríš, že by som sa nemal báť vstúpiť do sekty?“ alebo „Takže... báť sa sekty je sektárstvo? To mi znie hodne sektársky.“ Takže posledný faktor mrzutosti – číslo sedem, ak počítate so mnou – je, že toto všetko sa tak dlho vysvetľuje.)

* →

→ http://lesswrong.com/lw/md/cultish_countercultishness/

K: Zanechávanie

121. Dôležitosť hovorenia: „joj!“

Práve som dočítal príbeh o páde Enronu, *Najbystrejší chlapi v miestnosti*, ktorý týmto vyhráva moje ocenenie za „Najneprimeranejší názov knihy“.

Neprekvapujúcou črtou pomalého rozkladu a náhleho zrútenia Enronu bolo, že výkonní hráči si nikdy nepriznali, že urobili *veľkú* chybu. Keď katastrofa číslo 247 narástla do takých rozmerov, že si vyžadovala skutočnú zmenu pravidiel, povedali si: „Škoda, že to nevyšlo – bol to taký dobrý nápad – ako teraz skryjeme tento problém v našej súvahe?“ Namiesto: „Zo spätného pohľadu sa teraz zdá jasné, že to od samého začiatku bola chyba.“ Namiesto: „Bol som hlúpy.“ Nikdy nenastala chvíľa prelievať slzy, chvíľa na ponižujúce uvedomenie, na priznanie si *základného* problému. Po bankrote, Jeff Skilling, bývalý prevádzkový riaditeľ a nakrátko výkonný riaditeľ Enronu, odmietol radu svojich právnikov využiť piaty dodatok ústavy; svedčil pred kongresom, že Enron bola *výborná* spoločnosť.

Nie každá zmena je zlepšenie, ale každé zlepšenie je nevyhnutne zmena. Ak si priznávame iba malé lokálne chyby, budeme robiť iba malé lokálne zmeny. Motivácia na *veľkú* zmenu vychádza z uvedomenia si *veľkej* chyby.

Ako dieťa som bol vychovaný rovnakým dielom vedou a vedeckou fantastikou, a od Heinleina a Feynmana som sa naučil postupy tradičnej rozumnosti: Teórie musia byť odvážne a vystavovať sa falzifikácii; buď ochotný podstúpiť hrdinskú obeť vzdania sa svojich názorov, keď narazíš na dôkazy o opaku; argumentuj poctivo; snaž sa neklamať sám seba; a ďalšie nejasné slovné spojenia.

Výchova k tradičnej rozumnosti sa pokúša vytvoriť diskutérov, ktorí *jedného dňa* pripustia dôkaz o opaku – mala by existovať *nejaká* hora dôkazov dosť veľká na to, aby vami pohla. To nie je maličkosť; týmto sa odlišuje veda od náboženstva. Ale nie je tu veľký dôkaz na *rýchlosť*; na vzdanie boja *tak rýchlo, ako sa len dá*; na *efektívne* spájanie indícií, aby *minimum* indícií proti stačilo na zničenie vášho obľúbeného názoru.

Bol som vychovaný v tradičnej rozumnosti a považoval som sa za dosť rozumného. Na bayesovské remeslo (Laplace / Jaynes / Tversky / Kahneman) som sa prepol až v dôsledku... no, to by bol dlhý príbeh. Približne povedané, prepol som sa, keď som si uvedomil, že nejasné slovné spojenia tradičnej rozumnosti mi nestačili zabrániť v urobení veľkej chyby.

Po tom, čo som si konečne a naplno uvedomil svoju chybu, obzrel som sa na cestu, ktorá ma priviedla k môjmu strašnému uvedomeniu. A videl som, že som robil série malých ústupkov, minimálnych priznaní, neochotných ústupkov o každý milimeter územia, pri každej príležitosti si uvedomujúc najmenšie možné množstvo vlastných chýb, priznávajúc porážku iba v malých znesiteľných sústach. Mohol som sa hýbať omnoho rýchlejšie, uvedomil som si, keby som jednoducho zvrieskol: *„JOJ!“*

A tak som si pomyslel: *Musím sa túto hru naučiť hrať lepšie.*

V priznaní si *veľkej* chyby sa skrýva jedna *mocná výhoda*. Bolí to. Môže to aj zmeniť celý váš život.

Je *dôležité* mať chvíľu na prelievanie sĺz, chvíľu na ponižujúce uvedomenie. Priznať si *základný* problém, nerozdeľovať ho na stráviteľné sústa drobných chýb.

Neoddávajte sa dráme a nestaňte sa <u>hrdými na priznávanie chýb</u>. Iste je lepšie urobiť to správne na prvýkrát. Ale ak už urobíte chybu, potom je najlepšie uvidieť ju celú naraz. Ešte aj z hľadiska príjemnosti, lepšie je prijať jednu veľkú stratu než veľa malých strát. Alternatíva je naťahovať súboj so sebou samým po celé roky. Alternatíva je Enron.

Odvtedy som videl druhých, ako robia svoje vlastné série minimálnych ústupkov, neochotne ustupujúc o každý milimeter územia; nikdy si nepriznajúc globálnu chybu, ak stačí lokálna; vždy sa učiac z každej chyby tak málo, ako sa len dá. Čo by mohli napraviť jedným šmahom, keby sa im chcelo, premenia na maličké lokálne záplaty, do ktorých ich treba dotlačiť. Nikdy nepovedia po priznaní si chyby: *bol som blázon*. Robia, čo je v ich silách, aby minimalizovali svoju hanbu slovami: *v princípe som mal pravdu*, alebo *mohlo to fungovať*, alebo *stále sa chcem držať pravej podstaty toho-k-čomu-som-*

pripútaný. V danej chvíli bránia svoju hrdosť, čím si zabezpečujú, že danú chybu opäť zopakujú, a opäť budú musieť brániť svoju hrdosť.

To radšej prehltnúť celú horkú pilulku na jeden strašný glg.

⁂→

122. Ponuka pomätenosti

Keď som bol veľmi mladý – tuším trinásť- alebo štrnásťročný – myslel som si, že som našiel vyvrátenie Kantorovho diagonálneho argumentu, známej vety, ktorá ukazuje, že reálnych čísel je viac než racionálnych. Ach, tie predstavy slávy a úspechu, ktoré mi tancovali v hlave!

Moja myšlienka bola, že ak môžeme každé celé číslo rozložiť na súčet mocnín dvojky, môžeme namapovať prirodzené čísla na množinu podmnožín prirodzených čísel jednoducho tým, že vypíšeme ich dvojkový zápis. Napríklad 13, v dvojkovej sústave 1101, namapujeme na {0, 2, 3}. Trvalo celý týždeň než mi napadlo, že by som mohol skúsiť *aplikovať* Kantorov diagonálny argument na moju chytrú konštrukciu, a samozrejme sa našiel protipríklad – dvojkové číslo ...1111, ktoré nezodpovedá žiadnemu konečnému celému číslu.

Takže som našiel tento protipríklad a videl, že môj pokus o vyvrátenie bol falošný, rovnako ako moje sny o sláve a úspechu.

Spočiatku som bol trochu sklamaný.

Mysľou mi prešla myšlienka: „Raz tú vetu dostanem! *Jedného dňa* Kantorov diagonálny argument vyvrátim, aj keď mi to na prvýkrát nevyšlo!“ Neznášal som tú vetu, že je tak tvrdohlavo pravdivá, že ma obrala o moju slávu a úspech, a začal som hľadať iné vyvrátenie.

A potom som si čosi uvedomil. Uvedomil som si, že som urobil chybu a že teraz, keď som si svoju chybu všimol, nie je absolútne žiaden dôvod podozrievať silu Kantorovho diagonálneho argumentu, o nič viac než ľubovoľnej inej významnej matematickej vety.

Vtedy som uvidel veľmi jasne, že sa mi ponúkala príležitosť stať sa matematickým pomätencom a stráviť zvyšok svojho života písaním zeleným atramentom zlostných listov pre matematických profesorov. (Raz som čítal knihu o matematických pomätencoch.)

Nechcem som, aby toto bola moja budúcnosť, tak som sa krátko zasmial a nechal to tak. Zamával som Kantorovmu diagonálnemu argumentu na rozlúčku so želaním všetkého dobrého a už som ho viac nespochybňoval.

A teraz si nepamätám, či som si pomyslel už vtedy alebo či som si pomyslel až dodatočne... aký to bol hrozne neférový test pre trinásťročné dieťa. Že som musel už v tom veku byť rozumný alebo zlyhať.

Čím chytrejší ste, tým mladší môžete byť, keď po prvýkrát dostanete niečo, čo vyzerá ako naozaj revolučná myšlienka. Mal som šťastie, že som si tú chybu našiel sám; že mi ju nemusel ukázať nejaký iný matematik, a že mi tým nedal vonkajší predmet obviňovania. Možno by som sa aj tak jedného dňa zotavil. Zotavil som sa aj z horších vecí, ako dospelý. Ale keby som sa bol pomýlil tak skoro, vyvinul by som si vôbec túto schopnosť?

Rozmýšľam, koľkí ľudia píšuci zlostné listy zeleným atramentom mali trinásť rokov, keď urobili svoj prvý osudový prešľap. Rozmýšľam, koľkí z nich boli predtým nádejnými mysľami.

Urobil som chybu. To bolo všetko. Nemal som *v skutočnosti, v hlbšom zmysle pravdu*; nedosiahol som morálne víťazstvo; nepreukázal som ambíciu ani skepticizmus ani žiadnu inú úžasnú cnosť; nebola to rozumná chyba; nemal som ani spolovice pravdu ani najmenší zlomok pravdy. Myslel som si myšlienku, ktorú by som si nebol myslel, keby som bol múdrejší, a to je všetko, čo sa vtedy stalo.

Keby som si toto nebol schopný priznať, keby som bol reinterpretoval svoju chybu ako cnostnú, keby som bol trval na tom, že som mal aspoň *čiastočnú* pravdu, aby som si zachoval hrdosť, potom by som sa nebol nevzdal. Naďalej by som hľadal chybu v diagonálnom argumente. A skôr či neskôr by som možno nejakú našiel.

→ http://lesswrong.com/lw/i9/the_importance_of_saying_oops/

Dokiaľ si nepriznáte, že ste sa mýlili, nemôžete pokračovať vo svojom živote; váš sebaobraz bude stále pripútaný k tej starej chybe.

Kedykoľvek cítite pokušenie držať sa myšlienky, ktorú by ste si nikdy neboli pomysleli, keby ste boli múdrejší, ponúka sa vám príležitosť stať sa pomätencom – dokonca aj keď nikdy nenapíšete žiaden zlostný list zeleným atramentom. Ak sa s vami nikto neunúva hádať, alebo ak nikdy nikomu svoju myšlienku nepoviete, stále môžete byť pomätencom. *Lipnutie* je tá definujúca časť.

Nie je to pravda. Nie je to pravda v hlbšom zmysle. Nie je to ani polopravda, ani trochu pravda. Nie je to nič len myšlienka, ktorú ste si nikdy nemali pomyslieť. Nie na konci každého tunela je svetlo. Ľudia robia chyby a nie každá z nich je maskovaný úspech. Ľudia robia chyby; to sa stáva, a to je celé. Povedzte „joj“ a pokračujte vo svojom živote.

⁂→

123. Už to konečne vzdaj

Casey Serin, 24-ročný programátor webových stránok s nulovými skúsenosťami v obchode s nehnuteľnosťami dlží bankám 2,2 milióna dolárov po tom, čo klamal na žiadostiach o hypotéku, aby si mohol naraz kúpiť 8 rôznych domov v rôznych štátoch. Časť peňazí, ktoré získal ako hypotéku, vybral v hotovosti (žiadal o väčšie sumy než domy naozaj stáli) a minul ich na svoje životné náklady a semináre o obchodovaní s nehnuteľnosťami. Zdá sa, že očakával rast cien na trhu.

To ešte nie je tá smutná časť. Smutná časť je, že *to ešte stále nevzdal*. Casey Serin porážku nepripúšťa. Odmieta vyhlásiť bankrot, alebo si nájsť prácu; stále si myslí, že dokáže v obchode s nehnuteľnosťami preraziť. Naďalej míňa peniaze na semináre. Pokúsil sa získať hypotéku na deviaty dom. Ako vidíte, nebol to *neúspech*, bola to iba *príležitosť niečo sa naučiť*.

Toto sa stáva, keď sa odmietnete vzdať.

Zatiaľ čo toho správanie môže vyzerať ako obyčajná hlúposť, pripomína mi to dvoch ekonómov, nositeľov Nobelovej ceny...

...menovite Mertona a Scholesa z firmy Long-Term Capital Management.

Počas prvých troch rokov svojej existencie mohlo LTCM prehadzovať zisky lopatou, v roku 1998 sa však neefektivity, z ktorých LTCM ťažilo, začali strácať – viacerí ľudia poznali daný trik, preto prestal fungovať.

LTCM sa odmietlo vzdať. Navyknutí na ročné výnosy 40 %, požičiavali si na väčšie a väčšie páky, aby vyžmýkali drobnejšie a drobnejšie rozdiely. Keď začalo ísť pre LTCM všetko dole kopcom, mali kapitál 4,72 miliardy dolárov, páku 124,5 miliardy dolárov, a derivátové pozície za 1,25 bilióna.

Každé povolanie má iný spôsob, ako byť chytrý – treba sa naučiť iné zručnosti a riadiť sa inými pravidlami. Mohli by ste si preto myslieť, že štúdium „rozumnosti“ ako všeobecnej disciplíny nebude príliš prispievať k úspechu v skutočnom živote. Mne sa však aj tak zdá, že *ako nebyť hlúpy*, v tom majú viaceré povolania veľa spoločného. Ak sa pustíte do vyučovania *ako nenechať malé chyby prerásť do veľkých*, je to takmer to isté umenie, či už v hedžových fondoch alebo v láske, a jedna dôležitá vec je: Buď pripravený pripustiť, že si prehral.

⁂→

124. Správne použitie pochybnosti

Raz, keď som rečnil o Ceste, poznamenal som, že väčšina organizovaných systémov viery existuje na *únik pred pochybnosťou*. Jeden poslucháč zareagoval, že na jezuitov sa táto kritika nevzťahuje, pretože oni cvičia organizovanú pochybnosť: ich novicom sa vraj hovorí, aby pochybovali o kresťanstve, pochybovali o existencii Boha, pochybovali o pravosti svojho povolania, pochybovali o svojej vhodnosti

→ http://lesswrong.com/lw/j8/the_crackpot_offer/

→ http://lesswrong.com/lw/gx/just_lose_hope_already/

na večný sľub čistoty a chudoby. Povedal som: *Aha, ale predpokladá sa, že tieto pochybnosti prekonajú, nie?* On povedal: *Nie, majú pochybovať asi preto, aby ich pochybnosti mohli rásť a silnieť.*

Google mi tieto tvrdenia ani nepotvrdil ani nevyvrátil. (Ak mi niekto z vás vie pomôcť, bol by som mu zaviazaný.) Ale mne takýto scenár pripadá fascinujúci, hodný diskusie, bez ohľadu na to, či to je alebo nie je pravda o jezuitoch. *Keby* jezuiti cvičili úmyselné pochybovanie, ako je opísané hore, boli by *preto* cnostnými racionalistami?[142]

Myslím, že by som musel pripustiť, že jezuitov v horeuvedenom (údajnom) scenári nie je primerané opísať ako „unikajúcich pred pochybnosťou". Ale aj toto (údajné) správanie sa mi zdá vysoko podozrivé. Pre naozaj cnostného racionalistu by pochybnosť nemala byť strašidelná. Horepopísané správanie mi znie ako program na desenzitizáciu na niečo *veľmi* strašidelné, ako keď arachnofóba vystavujeme pavúkom za starostlivo kontrolovaných podmienok.

Ale aj tak, povzbudzujú svojich novicov, aby pochybovali – správne? Záleží na tom, či sú ich dôvody pomýlené? Nie je to napriek tomu konanie hodné racionalistu?

Zvedavosť sa snaží zničiť sama seba; niet takej zvedavosti, ktorá by *nechcela* odpoveď. Ale ak získate odpoveď, ak uspokojíte svoju zvedavosť, potom to slávne tajomstvo už viac nebude tajomným.

Rovnako, každá pochybnosť existuje preto, aby zničila nejaký konkrétny názor. Ak sa pochybnosti nepodarí zničiť svoj cieľ, zomiera nenaplnená – ale aj to je nejaké riešenie, nejaké ukončenie, aj keď smutnejšie. Pochybnosť, ktorá nezničí ani sama seba ani svoj cieľ, akoby ani nikdy neexistovala. *Riešenie* pochybnosti, nie samotný akt pochybovania, je to, čo otáča ozubené koleso rozumnosti vpred.

Každé zlepšenie je zmena, ale nie každá zmena je zlepšenie. Každý racionalista pochybuje, ale nie každá pochybnosť je rozumná. Nosenie pochybností vás nerobí racionalistom o nič viac než nosenie bieleho laboratórneho plášťa lekárom.

Rozumná pochybnosť vzniká z konkrétneho dôvodu – máte nejaký konkrétny dôvod pochybovať, že podozrivý názor je pravdivý. Tento dôvod potom vyžaduje nejakú líniu vyšetrovania, ktorá buď zničí cieľový názor alebo zničí túto pochybnosť. To platí aj pre vysoko abstraktné pochybnosti ako: „Ktovie, či neexistuje jednoduchšia hypotéza, ktorá by tiež vysvetlila tieto údaje." V tomto prípade vyšetrujete tak, že sa pokúšate vymyslieť jednoduchšie hypotézy. Ako toto hľadanie pokračuje dlhšie a dlhšie bez výsledkov, myslíte si, že je menej a menej pravdepodobné, že nasledujúci krok výpočtu bude tým úspešným. Nakoniec cena hľadania prekročí očakávaný úžitok, a vy prestanete hľadať. V tom bude už viac nemôžete tvrdiť, že *užitočne pochybujete*. Pochybnosť, ktorá nevyšetruje, akoby ani neexistovala. Každá pochybnosť existuje preto, aby zničila sama seba, jedným či druhým spôsobom. Nevyriešená pochybnosť je nulová operácia; neotáča kolesom, ani dopredu, ani dozadu.

Keby ste naozaj verili náboženstvu (a nie iba verili v náboženstvo), prečo by ste potom hovorili svojim novicom, aby zvažovali pochybnosti, ktoré musia zomrieť nenaplnené? Bolo by to ako hovoriť študentom fyziky, aby do úmoru pochybovali, že revolúcia v 20. storočí mohla byť chybou, a že by newtonovská mechanika mohla byť stále správna. Ak o niečom *naozaj* nepochybujete, prečo by ste *predstierali* že áno?

Pretože všetci chceme vyzerať ako rozumní – a pochybovanie sa *všeobecne považuje* za cnosť racionalistu. Nie je však všeobecne jasné, že na pochybovanie potrebujete konkrétny dôvod, alebo že nevyriešená pochybnosť je nulová operácia. Namiesto toho si ľudia myslia, že je to o *skromnosti*; submisívne správanie, ktoré udržiava hierarchiu postavenia v tlupe – takmer rovnaký problém ako s pokorou, o čom som už písal. Urobiť veľkú verejnú ukážku pochybnosti, aby ste presvedčili sami seba, že ste rozumný, vám je užitočné asi rovnako ako nosiť laboratórny plášť.

Aby ste sa vyhli vyznávaniu pochybností, pamätajte, že:

- Rozumná pochybnosť existuje preto, aby zničila svoj cieľový názor, a ak svoj cieľový názor nezničí, zomiera nenaplnená.

142 Poznámka prekladateľa: V diskusii pod pôvodným článkom sa táto otázka celkom nevyriešila, ale zdá sa, že pochybovať o všetkom nie je presná formulácia. Od novicov sa vyžadujú dva roky podrobného sebaskúmania, aby vstúpili iba tí, ktorí to myslia vážne.

- Rozumná pochybnosť vzniká z nejakého konkrétneho dôvodu, prečo by daný názor mohol byť nesprávny.
- Nevyriešená pochybnosť je nulová operácia.
- Nevyšetrená pochybnosť akoby ani neexistovala.
- Nemali by ste byť hrdí na púhe pochybovanie, môžete však byť právom hrdí, keď ste akurát *dokončili* trhanie svojho obľúbeného názoru na kúsky.
- Aj keď si môže vyžadovať veľa odvahy čeliť svojim pochybnostiam, nezabúdajte nikdy na to, že *ideálna myseľ* by sa pochybnosti v prvom rade nikdy nebála.

*→

125. Dokážete čeliť skutočnosti

Čo je pravda, to už je pravda.

Priznať si to to nerobí horším.

Nepriznať si to nespôsobí, že to zmizne.

A pretože je to pravda, je to to, s čím sa tu dá interagovať.

Čokoľvek nepravdivé tu nie je, nezažijeme to.

Ľudia dokážu zniesť to, čo je pravda,

pretože už to znášajú.

--Eugene Gendlin

*→

126. Meditácia o zvedavosti

Prvou cnosťou je zvedavosť

--Dvanásť cností rozumnosti

Ako racionalisti, sme povinní kritizovať sami seba a spochybňovať svoje názory... alebo nie?

Uvedomte si, čo sa s vami stane na psychologickej úrovni, keď začnete slovami: „Je mojou povinnosťou kritizovať svoje vlastné názory.“ Roger Zelazny raz ukázal na rozdiel medzi: „chcieť byť spisovateľom“ a „chcieť písať“. Mark Twain povedal: „Klasika je niečo, čo by každý chcel mať prečítané, ale nikto to nechce čítať.“ Kritizovať sám seba z pocitu povinnosti znamená *chcieť mať preskúmané*, aby ste potom mohli povedať, že vaša viera nie je slepá. To nie je to isté ako *chcieť skúmať*.

Môže to viesť k motivovanému zastaveniu vášho skúmania. Zvážite námietku, potom protiargument na túto námietku, a *tam sa zastavíte*. Toto zopakujete s niekoľkými námietkami, dokiaľ nemáte pocit, že ste splnili svoju povinnosť skúmať, a *tam sa zastavíte*. Dosiahli ste svoj základný psychologický cieľ: zbaviť sa kognitívnej disonancie, ktorá by vznikla z toho, že sa považujete za racionalistu a predsa viete, že ste sa nepokúsili kritizovať svoje názory. Mohli by sme to nazvať nakupovaním racionalistického uspokojenia – snaha vytvoriť „hrejivý pocit“ splnenej povinnosti.

Potom bude vaša uvádzaná úroveň pravdepodobnosti dosť vysoká na to, aby zdôvodnila, prečo sa držíte svojich pôvodných plánov a názorov, ale nie taká vysoká, aby vyvolala nedôverčivosť vašu alebo iných racionalistov.

Keď ste naozaj zvedaví, budú vás priťahovať otázky, ktoré vyzerajú najsľubnejšie na spôsobenie zmeny názoru, alebo otázky, ktoré sa najmenej podobajú tým, ktoré ste skúšali predtým. Na konci vaša

→ http://lesswrong.com/lw/ib/the_proper_use_of_doubt/

→ http://lesswrong.com/lw/id/you_can_face_reality/

pravdepodobnostná distribúcia asi *nebude* vyzerať tak, ako keď ste začali – nastanú zmeny, či nahor alebo nadol; ľubovoľný smer vám je rovnako dobrý, ak ste úprimne zvedaví.

Porovnajte si toto s podvedomým motívom držať svoje otázky na známom území, aby ste mohli svoje skúmanie rýchlo vybaviť, aby ste už *mali preskúmané*, a obnovili tým známu rovnováhu, na ktorej sú založené vaše staré známe plány a názory.

Ohľadom toho, ako si myslím, že by mala vyzerať pravá zvedavosť, a akú má silu, odporúčam vám Bájku o vede a politike. Každý z týchto postáv má vykresľovať inú lekciu. Ferris, posledná postava, stelesňuje silu nevinnej zvedavosti: čo je ľahkosť a dychtivé načahovanie sa po indíciách.

Ursula K. LeGuin napísala: „Nevinnosť nemá silu proti zlu. Ale má silu pre dobro.“[143] Nevinná zvedavosť môže sa môže nevinne zvrtnúť; a preto treba výcviku racionalistu a s ním súvisiacej sofistikovanosti čeliť ako nebezpečenstvu, ak sa chceme stať silnejšími. Napriek tomu sa môžeme pokúšať udržať si túto ľahkosť a dychtivé načahovanie nevinnosti.

Ako je napísané v Dvanástich cnostiach:

> Ak vo svojom srdci veríš, že už vieš, alebo ak si vo svojom srdci neželáš vedieť, potom tvoje vypytovanie bude bez cieľa a tvoje zručnosti bez smeru. Zvedavosť sa snaží zničiť sama seba; neexistuje zvedavosť, ktorá by netúžila po odpovedi.

Neexistuje skrátka žiadna dobrá náhrada za úprimnú zvedavosť. „Pálčivé svrbenie poznať je viac než slávnostná prísaha hľadať pravdu.“ Nemôžete však vytvoriť zvedavosť silou vôle o nič viac než silou vôle donútiť svoje nohy, aby cítili teplo, keď cítia chlad. Občas jediné, čo máme, sú naše púhe slávnostné prísahy.

Čo teda môžete urobiť s povinnosťou? Na začiatok, môžeme sa skúšať zaujímať o naše povinné skúmania – pozorne striehnuť na iskry skutočného záujmu, alebo aspoň skutočnej nevedomosti a túžby vyriešiť ju. Toto ide spolu so striehnutím na možnosti, ktoré sú bolestivé, pred ktorými cúvate – nie všetko z toho je negatívne myslenie.

Malo by vám aj pomôcť meditovať o Zákone zachovania očakávanej indície. Pre každý *nový* bod skúmania, pre každý kúsok dosiaľ *nevidenej* indície, na ktorú náhle pozriete, očakávaná výsledná pravdepodobnosť by sa mala rovnať vašej pôvodnej pravdepodobnosti. V mikroprocesoch skúmania by váš názor mal byť rovnako pripravený posunúť sa oboma smermi. Nie každý bod musí stačiť na širokú zmenu – na posunutie názoru z pravdepodobnosti 70% na 30% – ale ak je váš terajší názor 70%, mali by ste byť rovnako pripravení znížiť ho na 69% ako zvýšiť ho na 71%. Nemali by ste si myslieť, že viete, ktorým smerom pôjdete (v priemere), lebo podľa zákonov teórie pravdepodobnosti, ak poznáte svoj cieľ, už tam ste. Ak viete skúmať poctivo, takže každý *nový* bod má naozaj rovnaký potenciál posunúť váš názor nahor alebo nadol, môže vám to pomôcť zostať zaujatý alebo priam zvedavý na mikroprocesy skúmania.

Ak argument, ktorý zvažujete, *nie je* nový, prečo mu vôbec venujete pozornosť? Je to to, kam by ste sa pozerali, keby ste boli úprimne zvedaví? Kritizujete podvedome svoj názor na jeho najsilnejších miestach, namiesto najslabších? Precvičujete si indície?

Ak si dokážete neprecvičovať už známe obrany, a dokážete znížiť svoj názor o jeden malý kúsok z každej novej indície, možno budete dokonca schopní vzdať sa svojho názoru úplne – uvedomiť si, z ktorého smeru vietor indícií veje proti vám.

Ďalší posilňujúci prostriedok na zvedavosť je to, čo som začal volať Tarskeho litánia, čo je v skutočnosti meta-litánia, ktoré sa špecializuje na každý prípad (tak je to vhodné). Napríklad, ak som v napätí z rozmýšľania, či zamknutá krabica obsahuje diamant, potom si namiesto myslenia na krásne dôsledky, ak krabica obsahuje diamant, môžem opakovať Tarskeho litániu:

> Ak táto krabica obsahuje diamant,
>
> Chcem veriť, že táto krabica obsahuje diamant;
>
> Ak táto krabica neobsahuje diamant,

143 Ursula K. Le Guin, *The Farthest Shore* (Saga Press, 2001).

Chcem veriť, že táto krabica neobsahuje diamant;

Kiež nie som pripútaný k názorom, ktoré možno nechcem.

Potom môžete meditovať o možnosti, že tam nie je diamant, a o následnej výhode, ktorú budete mať, ak budete veriť, že tam nie je diamant, a o následnej nevýhode, ktorú budete mať, ak budete veriť, že tam diamant je. Pozrite si aj Gendlinovu litániu.

Ak v sebe dokážete nájsť najmenší kúsok pravej neistoty, potom si ho strážte ako keď lesník udržiava táborák. Ak dokážete dosiahnuť, aby z neho vzbĺkol plameň zvedavosti, urobí vás ľahkým a dychtivým, a dá zmysel vašim otázkam a smer vašim schopnostiam.

*→

127. Nikto vám nemôže udeliť výnimku zo zákonov rozumnosti

Tradičná Rozumnosť je formulovaná v pojmoch *spoločenských pravidiel*, ktorých porušovanie sa interpretuje ako podvádzanie – ako nedodržiavanie noriem spolupráce. Ak chcete, aby som prijal váš názor, ste povinný poskytnúť mi isté množstvo indície. Ak sa z toho skúšate vykrútiť, všetci vieme, že nedodržiavate svoje povinnosti. Teória je povinná robiť vlastné odvážne predpovede, nielen kradnúť predpovede, ktoré si odmakali iné teórie. Teória je povinná vystaviť sa falzifikácii – ak sa pokúša vykrútiť, je to ako pokúšať sa vykrútiť z obávaného iniciačného rituálu; musíte zaplatiť svoje dlhy.

Tradičná Rozumnosť je formulovaná podobne ako zvyky, ktorými sa riadia ľudské spoločnosti, čo zjednodušuje jej ústne odovzdávanie. Ľudia si všímajú spoločenské podvody omnoho spoľahlivejšie než izomorfné porušenia abstraktných logických pravidiel. Vnímanie rozumnosti ako spoločenskej povinnosti však prináša niektoré zvláštne myšlienky.

Napríklad človek vidí, ako veriaci ľudia obraňujú svoju vieru slovami: „*Ty* zase nevieš odôvodniť, prečo veríš vo vedu!“ Inými slovami: „Ako sa opovažuješ kritizovať ma za neodôvodnené názory, pokrytec! Ty to tiež robíš!“

Pre Bayesovca je mozog nástrojom presnosti: spracováva a sústreďuje previazané indície do mapy, ktorá odráža územie. Princípy rozumnosti sú zákonmi v rovnakom zmysle ako druhý zákon termodynamiky: získať spoľahlivý názor si vyžaduje vypočítateľné množstvo previazanej indície, rovnako ako spoľahlivo ochladiť obsah chladničky si vyžaduje vypočítateľné minimum voľnej energie.

Zákony fyziky sú v princípe časovo reverzibilné, takže existuje nekonečne malá pravdepodobnosť – neodlíšiteľná od nuly pre všetkých okrem matematikov – že chladnička sa ochladí spontánne a ešte pritom vygeneruje elektrinu. Existuje o čosi väčšia nekonečne malá pravdepodobnosť, že by ste dokázali správne nakresliť podrobnú mapu ulíc New Yorku, hoci by ste ho nikdy nenavštívili, iba by ste sedeli vo svojej obývačke so zatiahnutými závesmi a bez internetového pripojenia. Ale ja by som sa na to nespoliehal.

Skôr než sa pokúsite mapovať nevidené územie, nalejte si trocha vody do šálky pri izbovej teplote a najprv čakajte, či spontánne nezamrzne. Iba vtedy si môžete byť istí, že váš trik – ignorovanie nekonečne malých pravdepodobností úspechu – funguje vo všeobecnosti správne. Možno si hneď neuvedomíte, že vaša mapa je nesprávna, najmä ak ste v New Yorku nikdy neboli; ale môžete vidieť, že voda sama od seba nemrzne.

Ak sú pravidlá rozumnosti spoločenskými zvykmi, potom sa zdá možným ospravedlniť správanie X ak poukážete na to, že druhí robia to isté. Nebolo by *férové* žiadať indície od vás, ak ich nedokážeme poskytnúť sami. Uvedomíme si, že nikto z nás nie je lepší než ostatní, ustúpime a láskavo vás ospravedlníme z vašej spoločenskej povinnosti poskytnúť indície pre vaše presvedčenie. A potom budeme všetci naveky žiť šťastne v slobode, bratstve a rovnosti.

Ak sú pravidlá rozumnosti matematickými zákonmi, potom bude snaha ospravedlniť presvedčenie bez indícií poukázaním na to, že niekto iný robí to isté, asi rovnako efektívna ako vymenovanie 30 dôvodov, prečo by ste nemali spadnúť z útesu. Aj keby sme všetci odhlasovali, že je neférové, že vaša

→ http://lesswrong.com/lw/jz/the_meditation_on_curiosity/

chladnička potrebuje elektrinu, aj tak nebude fungovať (s pravdepodobnosťou ~1). Aj keby sme všetci odhlasovali, že nemusíte navštíviť New York, mapa bude aj tak nesprávna. Pani Príroda je známa svojou ľahostajnosťou voči podobným prosbám, a takisto aj pani Matematika.

Takže – aby sme sa vrátili k spoločenskému jazyku Tradičnej Rozumnosti – nemyslite si, že *vám prejde* tvrdenie, že je okej mať svojvoľný názor na XYZ, pretože iní ľudia tiež majú svojvoľné názory. Ak sa dve strany zmluvy obe správajú rovnako nesprávne, ľudský sudca sa môže rozhodnúť neudeliť pokutu žiadnej z nich. Ale ak dvaja inžinieri navrhnú svoj stroje rovnako nesprávne, ani jeden stroj nebude fungovať. Jedna chyba v návrhu nemôže ospravedlniť druhú. Aj keď *ja* robím XYZ nesprávne, nepomôže vám to, ani vám to nedá výnimku z pravidiel; znamená to akurát, že sme obaja v kýbli.

Čo sa týka ľudského zákona v liberálnych demokraciách, každý má nárok na svoj názor. Čo sa týka zákona Prírody, nemáte nárok na presnosť. Nezatýkame ľudí za to, že si myslia divné veci, prinajmenšom v tých múdrejších krajinách nie. Nikto však nemôže zrušiť zákon, že potrebujete indície na vytvorenie *presných* názorov. Ani hlasovanie celej ľudskej rasy nedokáže získať zhovievavosť na súde Prírody.

Fyzici nerozhodujú o fyzikálnych zákonoch, oni len hádajú, aké sú. Racionalisti nerozhodujú o zákonoch rozumnosti, iba hádame, aké sú. Nemôžete „racionalizovať“ niečo, čo od začiatku nie je rozumné. Keby sa vám darom mimoriadnej presvedčivosti podarilo presvedčiť všetkých fyzikov na svete, že máte výnimku zo zákonov gravitácie, a skočíte z útesu, spadnete. Dokonca samotné slová „*my* nerozhodujeme“ sú príliš antropomorfné. Neexistuje žiadna vyššia moc, ktorá by vám mohla dať výnimku. Existuje iba príčina a následok.

Pamätajte na to, keď budete žiadať aspoň o jednorazovú výnimku. My vám ju *nemôžeme* dať. Od nás to nezávisí.

*→

128. Ponechajte ústupovú líniu

> Keď obkľúčiš nepriateľa,
>
> Vždy mu ponechaj únikovú cestu.
>
> Musí vidieť, že existuje
>
> Alternatíva k smrti.
>
> --Sun C', *Umenie vojny*[144]
>
> Nezvyšuj tlak, znižuj odpor.
>
> --Lois McMaster Bujold, *Komarr*[145]

Včera večer som sa rozprával s neracionalistkou, ktorá akosi zablúdila na zhromaždenie miestnych racionalistov. Vyhlásila, že (a) verí, že existujú duše, a (b) že neverí v kryoniku, lebo verí, že duša by nezostala v zmrazenom tele. Opýtal som sa: „Ale ako to vieš?“ Zo zmätku, ktorý jej prebleskol tvárou bolo celkom jasné, že jej táto otázka nikdy nenapadla. Nehovorím to v zlom – vyzerala ako milá osoba s absolútne žiadnym výcvikom v rozumnosti, tak ako väčšina zvyšku ľudského druhu. Naozaj musím tú knihu napísať.

Väčšina výslednej konverzácie bola na témy už pokryté na *Overcoming Bias* – ak vás niečo *naozaj* zaujíma, pravdepodobne *dokážete* vymyslieť dobrý spôsob, ako to otestovať; pokúste sa najprv dosiahnuť presné názory a až potom z nich nechajte vyplynúť svoje emócie – takéto veci. Ale tento rozhovor mi pripomenul jednu vec, ktorú som tu ešte nepreberal.

„Ubezpeč sa,“ navrhol som jej, „že si dokážeš predstaviť, ako by vyzeral svet, keby neexistovali žiadne duše, a čo by si vtedy robila. Nemysli na všetky dôvody, prečo to tak nemôže byť, jednoducho to

→ http://lesswrong.com/lw/k1/no_one_can_exempt_you_from_rationalitys_laws/

144 Sun Tzu, *The Art of War* (Cloud Hands, Inc., 2004).

145 Lois McMaster Bujold, *Komarr*, Miles Vorkosigan Adventures (Baen, 1999).

prijmi ako predpoklad a potom si predstav dôsledky. Takže si pomyslíš: ‚No, keby duše neexistovalo, mohla by som sa prihlásiť na kryoniku' alebo ‚Keby neexistoval Boh, aj tak by som mohla byť naďalej morálna' namiesto toho, aby to bola príliš strašná predstava. Ako vec sebaúcty by si sa mala pokúsiť veriť tomu, čo je pravda, bez ohľadu na to, aké je to nepohodlné, ako som už povedal; ale ako vec ľudskej povahy, pomáha, keď si ten názor urobíme menej nepohodlným, *skôr než* začneme vyhodnocovať indície v jeho prospech."

Princíp za touto technikou je jednoduchý: Ako radí Sun C' ohľadom vašich nepriateľov, tak musíte aj sami so sebou – ponechajte si ústupovú líniu, aby ste mali menej problémov s ustupovaním. Predstava, že prídete o svoju prácu, napríklad, môže vyzerať omnoho strašidelnejšie, keď na to ani nedokážete pomyslieť, než keď ste si prepočítali, ako dlho vám vydržia úspory, pozreli ste si trh práce vo vašej oblasti, a inak ste presne naplánovali, čo urobíte ďalej. Iba potom budete pripravení *poctivo* zhodnotiť pravdepodobnosť, že si udržíte svoju prácu po plánovanom prepúšťaní budúci mesiac. Buďte skutočným zbabelcom, a naplánujte si váš ústup do podrobností – predstavte si každý krok – najlepšie ešte skôr než prvýkrát vstúpite na bojisko.

Nádej je, že si vyžaduje menej odvahy predstaviť si nepohodlný stav vecí *ako myšlienkový experiment*, než uvažovať, *s akou pravdepodobnosťou* je to pravda. Ale potom, čo urobíte to prvé, je ľahšie urobiť to druhé.

Pamätajte, že bayesovstvo je presné – dokonca aj keď strašidelná predstava vyzerá nepravdepodobne, je stále dôležité spočítať všetky indície za a proti, presne férovo, aby ste došli k rozumnej kvantitatívnej pravdepodobnosti. Vizualizovať si strašidelnú predstavu *neznamená* priznať, že si hlboko vnútri myslíte, že je pravdivá. Môžete si predstavovať strašidelné názory v rámci dobrého myšlienkového upratovania. „Myšlienka, ktorú si nedokážte pomyslieť vás ovláda viac než myšlienky, ktoré vyslovujete nahlas" - to sa stane, ešte aj keď je tá nemysliteľná myšlienka nepravdivá!

Technika ponechania ústupovej línie si vyžaduje isté minimum sebapoctivosti, ak sa má používať správne.

Na začiatok: Musíte byť prinajmenšom schopní pripustiť si, *ktoré* myšlienky vás desia, a ku ktorým myšlienkam ste pripútaní. Ale toto je podstatne menej ťažké než férovo spočítať indície za myšlienku, ktorá vás desí. Pomôže vám, keď poviem, že som občas túto techniku sám použil? Racionalista predsa neodmieta všetky emócie. Existujú myšlienky, ktoré ma desia, ale aj tak verím, že sú nepravdivé. Existujú myšlienky, ku ktorým som pripútaný, ale predsa verím, že sú pravdivé. Ale aj tak stále plánujem svoje ústupy, nie preto, že by som plánoval *ustúpiť*, ale pretože plánovanie ústupu vopred mi pomáha rozmýšľať o danom probléme bez pripútanosti.

Ale väčší test sebaúprimnosti je *naozaj* prijať nepohodlný návrh ako predpoklad, a zistiť, ako by ste si s ním *naozaj* poradili. Keď čelíme nepohodlnej myšlienke, náš prvý impulz je myslieť na všetky dôvody, prečo to *vôbec nemôže* byť tak. A tak v sebe nájdete isté množstvo psychologického odporu, keď sa pokúsite predstaviť si, aký presne by bol svet, a čo by ste s tým robili, keby Môj Najdrahší Názor bol nepravdivý, alebo Moja Najväčšia Obava bola pravdivá.

Myslite na všetkých tých ľudí, ktorí hovoria, že bez Boha je morálna nemožná. (A áno, táto téma sa v rozhovoroch vyskytuje; neponúkam tu slameného panáka.) Keby si veriaci vedeli predstaviť svoju *skutočnú* reakciu na to, keby uverili ako fakt, že Boh neexistuje, uvedomili by si, že by vlastne nešli zabíjať bábätká. Mohli by si uvedomiť, že ateisti reagujú na neexistenciu Boha viacmenej rovnako ako by reagovali oni sami, keby tomu uverili. Hovorím to, aby som ukázal, že *je* značne ťažké predstaviť si, ako by *ste naozaj* reagovali, keby ste verili opaku niečoho, čomu teraz pevne veríte.

Plus je vždy protiintuitívne uvedomiť si, že áno, ľudia sa cez veci nejako prenesú. Čerství vozíčkari nie sú o šesť mesiacov takí smutní, ako to dnes očakávajú, atď. Môže byť rovnako protiintuitívne uvedomiť si, že keby sa strašidelná predstava ukázala ako pravdivá, nejako *by* ste sa s ňou vyrovnali. Vozíčkari sa vyrovnajú, aj vy by ste sa vyrovnali.

Pozrite si aj Gendlinovu litániu a Tarskeho litániu. Čo je pravda, už je pravda; priznať si to to neurobí horšie. Nemali by ste sa báť iba si *predstaviť* svet, ktorého sa bojíte. Ak je ten svet už skutočný, predstava to nijako nezhorší; a ak *nie je* skutočný, predstava neurobí žiadnu škodu. A pamätajte si, keď si

predstavujete, že ak tie strašné veci, ktoré si predstavujete, sú naozaj pravdivé – čo môžu byť! - potom by ste tomu veru chceli veriť, a aj to by ste si mali predstaviť; neveriť by vám nepomohlo.

Koľko veriacich ľudí by si udržalo svoju vieru v Boha, keby si vedeli *presne* predstaviť hypotetický svet, v ktorom by Boh nebol a oni sami by sa stali ateistami?

Ponechať ústupovú líniu je mocná technika, ale nie je ľahká. *Úprimná* predstava si nevyžaduje toľko námahy ako *priamo* priznať, že Boh neexistuje, ale aj tak si vyžaduje námahu.

*→

129. Kríza viery

> Nie je to skutočná kríza viery, pokiaľ by veci nemohli rovnako ľahko ísť ľubovoľným smerom.
>
> --Thor Shenkel

Mnohí na tomto svete si udržiavajú názory, ktorých chyby by odhalilo desaťročné dieťa, *keby* toto desaťročné dieťa počulo daný názor po prvýkrát. Teraz nehovorím o nejakých rafinovaných chybách. Pre nepripútanú myseľ by bolo detskou hračkou zrieknuť sa ich, keby sa skepticizmus desaťročného použil bez vyhýbania. Ako napísal Premise Checker: „Keby sme predstavu boha neboli zdedili až do veku vedy, iba mimoriadne divný človek by vymyslel takú predstavu a tvrdil, že to niečo vysvetľuje."

A predsa zruční vedeckí špecialisti, dokonca veľkí inovátori vo svojej oblasti, ešte aj v dnešnej dobe neuplatňujú svoj skepticizmus úspešne. Nositeľ Nobelovej ceny Robert Aumann, autor Aumannovej vety o súhlase, je ortodoxný žid: považujem za veľmi pravdepodobné, že Aumann musel v nejakom bode svojho života zapochybovať o svojej viere. A predsa sa mu nepodarilo pochybovať úspešne. Svoje názory meníme zriedkavejšie než si myslíme.

Toto by vás malo vydesiť do špiku kostí. Znamená to, že môžete byť svetovým vedcom *a* mať bayesovskú matematiku v malíčku, *a* predsa nedokážete odmietnuť názor, ktorého absurditu by videli čerstvé desaťročné oči. Ukazuje to tú nezraniteľnú obrannú polohu, ktorú si názor dokáže pre seba vytvoriť, ak dosť dlho hnisal vo vašej mysli.

Čo treba, aby sme porazili chybu, ktorá si okolo seba vybudovala pevnosť?

Nuž, v čase, keď *viete*, že je to chyba, už bola porazená. Dilema neznie: „Ako sa dokážem zbaviť môjho dávneho nesprávneho názoru X?" ale: „Ako môže vedieť, že môj dávny názor X je nesprávny?" Úprimnosť k sebe samému je najjemnejšia, keď si nie sme *istí*, ktorá cesta je tá správna. A tak vzniká otázka:

Ako dokážeme v sebe vytvoriť skutočnú krízu viery, ktorá by rovnako ľahko mohla dopadnúť ľubovoľným smerom?

Náboženstvo je pokus, ktorý si vieme všetci predstaviť. (Čitatelia, ktorí mali rodičov ateistov, zmeškali základnú životnú skúšku, a musia sa uspokojiť s úbohou náhradou myslenia na svojich veriacich priateľov.) Ale ak ste preťali všetky sympatie a vnímate veriacich ako zlých mutantov, potom si nedokážete predstaviť skutočné vnútorné skúšky, ktorým čelia. Nebudete si vedieť položiť otázku:

„Akú všeobecnú stratégiu by veriaci človek mohol nasledovať, aby unikol zo svojho náboženstva?"

Som si istý, že pri pohľade na túto výzvu už niektorí zo seba sypú zoznam štandardných ateistických argumentov - „Museli by si priznať, že neexistuje žiadna bayesovská indícia pre existenciu Boha", „Museli by si uvedomiť morálne vyhýbanie, ktorého sa dopúšťajú, aby ospravedlnili správanie Boha v Biblii", „Museli by sa naučiť používať Occamovu britvu..."

ZLE! ZLE, ZLE, ZLE! Tento druh opakovania, kde iba vychŕlite body, *ktoré ste si už dávno premysleli*, je *presne* ten druh myslenia, ktorý udržiava ľudí pri v ich súčasných náboženstvách. Ak zostávate v rámci svojich uložených myšlienok, ak váš mozog dopĺňa samozrejmú odpoveď tak rýchlo, že ju nedokážete vidieť originálne, určite nedokážete navodiť krízu viery.

→ http://lesswrong.com/lw/o4/leave_a_line_of_retreat/

Možno je to otázka toho, že dosť ľudí nečítalo „Gödel, Escher, Bach“ v dostatočne mladom veku, ale všimol som si, že veľká časť populácie – dokonca aj technicky zameraných ľudí – má problém nasledovať argument, ktorý ide takto meta. V moje pesimistickejšie dni pochybujem, či ťava má dva hrby.

Ešte aj keď sú na to vyslovene upozornení, niektorí ľudia asi *nedokážu nasledovať skok* z objektovej úrovne: „Použi Occamovu britvu! Musíš vidieť, že tvoj Boh je nadbytočný názor!“ na meta-úroveň: „Skús zabrániť svojej mysli dopĺňať vzor zvyčajným spôsobom!“ Pretože rovnako ako vaši racionalistickí priatelia hovoria o tom, že Occamova britva je dobrá vec, a rovnako ako vám Occamova britva okamžite príde na myseľ, rovnako kolektívne schválená náboženská odpoveď je: „Božie cesty sú tajomné a je domýšľavé predstavovať si, že ich dokážeme pochopiť.“ Takže ak si vy myslíte, že *všeobecná* stratégia hodná nasledovania je: „Použi Occamovu britvu“, je to ako keď veriaci hovorí, že všeobecná stratégia je mať vieru.

„Ale... ale Occamova britva je naozaj lepšia než viera! To nie je ako uprednostňovať inú chuť zmrzliny! Každý sa môže pozrieť na históriu, že occamovské uvažovanie bolo omnoho produktívnejšie než viera...“

Čo je všetko pravda. Ale pointa je inde. Pointa je, že keď toho hovoríte, chŕlite zo seba štandardné zdôvodnenie toho, čo už je vo vašej hlave. Skutočná výzva krízy viery je zvládnuť prípad, kde je možné, že naše štandardné závery sú *nesprávne*, aj naše štandardné zdôvodnenia sú *nesprávne*. Takže ak štandardným zdôvodnením pre X je „Occamova britva!“, a chcete mať krízu viery ohľadom X, mali by ste zapochybovať, čo Occamova britva naozaj odporúča X, či vaše chápanie Occamovej britvy je správne, a – ak chcete mať dostatočne hlboké pochybnosti – či jednoduchosť *je* tým druhom kritéria, ktoré v minulosti v tomto prípade dobre fungovalo, alebo by sa dalo rozumne *očakávať*, že bude fungovať, atď. Keby ste poradili religionistovi zapochybovať o tom, či „viera“ je dobré zdôvodnenie pre X, potom by ste mali sebe odporučiť vynaložiť rovnako silné úsilie spochybniť svoju vieru, že „Occamova britva“ je dobrým zdôvodnením pre X.

(Pomyslite na všetkých tých ľudí, ktorí nerozumejú minimálnej dĺžke popisu alebo formulácii Occamovej britvy pomocou Solomonoffovej indukcie, ktorí si myslia, že Occamova britva vylučuje hypotézu mnohých svetov, alebo hypotézu simulácie. Potrebovali by zapochybovať o ich formulácii Occamovej britvy a svojej predstave, prečo je jednoduchosť dobrá vec. Ktorékoľvek X ste si práve zdôvodnili slovami „Occamova britva!“, stavím sa, že to nie je na rovnakej úrovni occamovského zásahu do terča ako gravitácia.)

Ak je „Occamova britva!“ vaša zvyčajná odpoveď, vaša štandardná odpoveď, odpoveď ktorú dávajú všetci vaši priatelia – potom by ste mali zablokovať svoj mozog od okamžitého dopĺňania tohto vzoru, ak sa pokúšate navodiť skutočnú krízu viery.

Lepšie je myslieť na pravidlá ako: „Predstav si, čo by povedal skeptik – a potom si predstav, čo by odpovedal na tvoju reakciu – a potom si predstav, čo ďalšie by mohol povedať, na čo by sa ťažšie odpovedalo.“

Alebo: „Skús si myslieť tú myšlienku, ktorá najviac bolí.“

A nadovšetko, toto pravidlo:

„Vynaložrovnakú úroveň zúfalého úsilia, aké by bolo treba, aby veriaci odmietol svoje náboženstvo.“

Pretože ak sa *nepokúšate* takto silno, potom – napriek všetkému, čo si *myslíte* – vaša hlava môže byť stále naplnená vecami rovnako nezmyselnými a smiešnymi ako náboženstvo.

Bez kŕčovitého krútiaceho úsilia byť rozumný, bez toho druhu úsilia, ktoré by bolo treba na odvrhnutie náboženstva – ako by ste sa odvážili veriť v niečo, keď Robert Aumann verí v Boha?

Niekto (zabudol som, kto) raz konštatoval, že ľudia majú šancu odmietnuť svoju náboženskú vieru iba do istého veku. Potom už majú odpovede na všetky námietky, a je príliš neskoro. Tento druh existencie musíte prekonať. Toto je test vašej sily ako racionalistu, a je to veľmi ťažké; ale ak ním nedokážete prejsť, budete slabší ako desaťroční.

Ale opäť, keď už viete, že nejaký názor je nesprávny, už ste ho porazili. Nehovoríme tu teda o podniknutí zúfalého kŕčovitého úsilia o zvrátenie účinkov náboženskej výchovy, *potom* čo ste došli k záveru, že náboženstvo je nepravdivé. Hovoríme tu o zúfalom úsilí *zistiť*, či by ste mali odhodiť svoje reťaze, alebo si ich ponechať. Úprimnosť k sebe samému je najjemnejšia, keď *nevieme*, po ktorej ceste máme ísť – vtedy racionalizácie nie sú *jasnými* hriechmi.

Nie každá pochybnosť si vyžaduje vykonanie kompletnej Krízy Viery. Ale mali by ste o nej uvažovať, keď:

- Nejaký názor bol dlho vo vašej mysli;
- Je obklopený oblakom známych argumentov a vyvrátení;
- Utopili ste v ňom náklady (čas, peniaze, verejné vyhlásenia);
- Tento názor má emocionálne dôsledky (to samotné neznamená, že je nesprávny);
- Je všeobecne previazaný s vašou osobnosťou.

Žiadne z týchto varovných znamení nie je automatické vyvrátenie. Tieto vlastnosti vytvárajú okolo ohrozeného názoru rôzne nebezpečenstvá, a robia jeho vyvrátenie veľmi ťažkým, ak *je* nesprávny. Ale rovnako by platili pre vieru Richarda Dawkinsa v evolučnú biológiu, ako vo vieru pápeža v katolicizmus. To neznamená, že sa tu rozprávame o rôznych príchutiach zmrzliny. Iba neosvietení si myslia, že všetky silné presvedčenia sú na rovnakej úrovni bez ohľadu na indície, ktoré ich podporujú, len preto, že sú silné. Naším cieľom nie je mať plytké názory, ale mať mapu, ktorá odráža územie.

Zdôrazňujem to, samozrejme, aby ste si mohli priznať: „Môj názor má tieto varovné znamenia," bez toho, že by ste si museli povedať: „Môj názor je nesprávny."

Ale čo tieto varovné znamenia *označujú*, je názor, o ktorom *efektívne pochybovanie si bude vyžadovať viac než len bežné úsilie*. Aby, pokiaľ je naozaj nesprávny, ste ho naozaj odmietli. A kde nedokážete účinne pochybovať, ste slepí, pretože váš mozog bude tento názor držať bezpodmienečne. Keď sietnica vysiela rovnaké signály bez ohľadu na to, aké fotóny do nej vstupujú, nazývame takéto oko slepým.

Kedy by ste mali podstúpiť Krízu Viery?

Opäť, pomyslite na radu, ktorú by ste dali veriacemu: Ak zistíš, že sa vo vnútri cítiš trochu nestabilne, ale pokúšaš sa racionalizovať dôvody, prečo je názor stále solídny, potom by si pravdepodobne mal vykonať rituál Krízy Viery. Ak je ten názor tak solídne podložený ako gravitácia, nemusíte sa unúvať – ale pomyslite na všetkých tých veriacich, ktorí by zúfalo chceli dôjsť k záveru, že Boh je rovnako solídny ako gravitácia. Skúste si teda predstaviť, čo by nejaký skeptik mohol povedať na váš argument „solídny ako gravitácia". Určite je jedným z dôvodov, prečo môžete v kríze viery zlyhať, to, že si v prvom rade nikdy nesadnete a nezačnete pochybovať – že nikdy nepoviete: „Tu je niečo, na čo musím vynaložiť úsilie, aby som správne pochyboval."

Ak vaše myšlienky začnú byť takéto komplikované, mali by ste ísť na to a vyvolať Krízu Viery. Neskúšajte to robiť chaoticky, neskúšajte to, ak máte iba chvíľku voľného času. Neponáhľajte sa, aby ste to čo najrýchlejšie mali za sebou, aby ste mohli povedať: „Pochyboval som, tak ako bolo mojou povinnosťou." To by nefungovalo pre veriaceho a nefungovalo by to ani pre vás. Predchádzajúci deň si oddýchnite, nech ste v dobrom mentálnom stave. Vyhraďte si pár hodín bez prerušenia. Nájdite si tiché miesto, kde si môžete sadnúť. Zbavte svoju myseľ všetkých štandardných argumentov, skúste vnímať od začiatku. A vynaložte zúfalé úsilie, aby ste vytvorili skutočnú pochybnosť, ktorá by dokázala zničiť nepravdivý, ale *iba* nepravdivý, hlboko držaný názor.

Prvky techniky Kríza Viery, rozhádzané po mnohých článkoch:

- Vyhýbanie sa naozaj slabým miestam vašej viery – Jedným z prvých pokušení krízy viery je pochybovať o najsilnejších miestach svojej viery, aby ste si mohli precvičiť svoje dobré odpovede. Potrebujete vyhľadať tie najboľavejšie miesta, nie argumenty, ktoré je najpríjemnejšie zvažovať.
- Meditácia o zvedavosti – Roger Zelazny raz rozlíšil medzi „chcieť byť autorom" a „chcieť písať", a takisto existuje rozdiel medzi chcieť mať niečo preskúmané a chcieť skúmať. Nestačí povedať: „Je mojou povinnosťou kritizovať vlastné názory"; musíte byť zvedaví a iba neistota dokáže vytvoriť zvedavosť. Pamätajte na Zákon zachovania očakávanej indície; môže vám to

pomôcť Aktualizovať postupne: Pri každom *jednom* bode, ktorý zvažujete, a pri každom prvku nového argumentu a novej indície, by ste nemali očakávať, že sa vaše názory posunú (v priemere) viac jedným smerom než druhým – takto môžete byť naozaj zakaždým zvedaví, ktorým smerom to pôjde.

- Uložené myšlienky a Pirsigov Originálny pohľad; zabráňte štandardným myšlienkam, aby vtrhli a doplnili vzor.
- Gendlinova litánia a Tarskeho litánia: Ľudia dokážu zniesť to, čo je pravda, lebo už to znášajú. Ak je názor pravdivý, bude pre teba lepšie, keď mu budeš veriť, a ak je nepravdivý, bude pre teba lepšie ho odmietnuť. Veriacemu človeku by si odporučil, aby si skúsil naplno a hlboko predstaviť svet, v ktorom neexistuje žiaden Boh, a aby bez výhovoriek naplno pochopil, že *ak* Boh neexistuje, *tak* preňho bude lepšie veriť, že Boh neexistuje. Ak človek nedokáže toto prijať na hlbokej emocionálnej úrovni, nedokáže mať krízu viery. Mali by ste teda vynaložiť úprimné úsilie predstaviť si *alternatívu* svojho názoru, tak ako by najlepší a najväčší skeptici chceli, aby ste si ju predstavili. Pomyslite na úsilie, ktoré musí vynaložiť religionista, aby si predstavil, bez prekrútenia pre svoje vlastné pohodlie, ateistov svetonázor.
- Vynaložte mimoriadne úsilie: pojem *isshokenmei*, zúfalé kŕčovité úsilie byť rozumný, ktoré by bolo treba na prekročenie úrovne Roberta Aumanna a všetkých veľkých vedcov v histórii, ktorí nikdy nezanechali svoje náboženstvá.
- Heuristika pôvodu: Mali by ste byť mimoriadne podozrievaví, ak máte mnoho myšlienok zo zdroja, o ktorom teraz viete, že je nedôveryhodný, ale zhodou okolností sa zdá, že všetky tieto myšlienky predsa len boli pravdivé. (Napríklad človek pripustí, že Bibliu napísali ľudia, ale stále sa drží predstavy, že obsahuje neodmysliteľnú etickú múdrosť.)
- Dôležitosť hovorenia „Joj“ - naozaj to menej bolí prehltnúť celú horkú tabletku na jeden hrozný glg.
- Singlethink, opak doublethinku. Pozrite sa na myšlienky, od ktorých cúvate, ktoré sa objavujú v kútiku vaše mysle iba na chvíľku, než na ne odmietnete myslieť. Ak si uvedomíte, na čo nemyslíte, môžete na to myslieť.
- Afektívne špirály smrti a Odolajte šťastnej špirále smrti. Afektívne špirály smrti sú hlavné generátory nepravdivých názorov, ktoré vyžadujú Krízu Viery, aby ste sa ich striasli. Ale keďže afektívne špirály smrti môžu vzniknúť aj okolo skutočných vecí, ktoré sú naozaj pekné, nemusíte pripustiť, že váš názor je lož, skúste na každom kroku odolať efektu svätožiary – odmietnite falošnú chválu aj naozaj dobrých vecí. Debaty o pravidlách by nemali vyzerať jednostranne.
- Počkajte s navrhovaním riešení, dokiaľ problém nebude prediskutovaný tak dôkladne, ako sa len dá bez ich navrhovania; zdržte svoju myseľ od poznania, čo bude jej odpoveď; a pokúšajte sa o to päť minút než sa vzdáte, aj vo všeobecnosti, ale najmä keď hľadáte diablov uhol pohľadu.

A tieto štandardné techniky sú mimoriadne relevantné:

- Časť o Spodnom riadku a Racionalizácii, ktorá vysvetľuje, prečo je vždy nesprávne selektívne argumentovať za jednu stranu debaty.
- Pozitívne skreslenie a motivovaný skepticizmus a motivované zastavenie, aby ste selektívne nehľadali podporu, selektívne nehľadali protiargumenty na protiargumenty, a selektívne nezastavili argument skôr než sa stane nebezpečným. Nevšímanie si alternatív je špeciálnym prípadom zastavenia sa. Špeciálnym prípadom motivovaného skepticizmu je falošná skromnosť, kde hanblivo priznáte, že nikto nemôže vedieť to, čo vy nechcete vedieť. Nepožadujte selektívne priveľa autority od protiargumentov.
- Pozor na Sémantické stopky, Signály na potlesk, a výber Vysvetli/Uctievaj/Ignoruj.
- Cíťte váhu Priťažujúcich podrobností; každá podrobnosť je samostatné bremeno, bod krízy.

Tu na *Overcoming Bias* je naozaj kopec relevantného materiálu. Kríza Viery je iba kritický bod a náhly náraz dlhodobého *isshoukenmei* – celoživotného nekompromisného úsilia byť taký neuveriteľne rozumný, že stúpnete nad úroveň hlúpych prekliatych chýb. Je to, keď dostanete šancu použiť svoje zručnosti, ktoré ste tak dlho precvičovali, naplno proti sebe samému.

Želám vám veľa šťastia proti vášmu súperovi. Majte nádhernú krízu!

*→

—

130. Rituál

Miestnosť, v ktorej Jeffreyssai prijímal svojich návštevníkov, ktorí neboli *beisutsukai*, bola ticho formálna, bezchybne určená iba tým najkonzervatívnejším vkusom. Slnečné svetlo a vonkajší vzduch prúdili cez mrežu z vyleštěného striebra, ktorej pár ostrých hrotov dávalo najavo, že táto stena sa nemá otvárať. Podlaha a steny boli zo skla, dosť hrubého na to, aby skresľovalo do tej miery, že nezáležalo na tom, čo by mohlo byť pod ním. Na povrchu skla boli jemne vyryté vzory bez konkrétneho významu, akoby načmárané rukou dieťaťa s umeleckými sklonmi (čo mimochodom bolo naozaj tak).

Inde v Jeffreyssaiovom dome boli miestnosti v inom štýle; ale ako zistil, toto bolo to, čo väčšina laikov očakávala od Bayesovského Majstra, a on sa rozhodol neosvecovať ich inak. Táto tichá zábava bola napokon jedným z drobných životných potešení.

Návšteva sedela oproti nemu, kolená na vankúši, podpätky za sebou. Bola tu výhradne kvôli záležitosti jej Konšpirácie, a jej odev to ukazoval: Priliehavá kombinéza z ružovej kože zakrývala aj jej ruky – až po kapucňu, ktoré jej prikrývala hlavu a vlasy, hoci tvár bola priama a nezakrytá.

A tak sa Jeffreyssai rozhodol, že ju prijme v tejto izbe.

Jeffreyssai zhlboka vydýchol. „Si si *istá*?“

„Ach,“ povedala, „a to si musím byť *absolútne istá* skôr než moja rada dokáže pohnúť tvojimi názormi? Nestačí ti, že som v danej oblasti odborníčka a ty nie?“

Jeffreyssaiove ústa sa v kútiku zdvihli v polovičnom úsmeve. „Odkiaľ *ty* vôbec vieš tak veľa o pravidlách? Nikdy si nemala ani len Planckovu dĺžku formálneho výcviku.“

„Je vôbec treba sa na to pýtať?“ povedala sucho. „Ak je niečo, o čom vy *beisutsukai* milujete rozprávať donekonečna, sú to dôvody, prečo robíte veci.“

Jeffreyssai sa vnútorne striasol pri predstave, že sa niekto pokúša pochytiť rozumnosť tým, že sleduje druhých, ako o nej hovoria...

„A nestriasaj sa vnútorne, keď ma počúvaš,“ povedala. „Ja sa nepokúšam byť racionalistka, iba sa s racionalistom snažím vyhrať debatu. V tom je rozdiel, ako určite rozprávaš svojim žiakom.“

Dokáže ma naozaj tak dobre čítať? Jeffreyssai sa pozrel von cez striebornú mrežu, a slnečné svetlo sa odrazilo od plochy hôr. Vždy, vždy zlaté slnečné svetlo zapadalo každý deň na tomto mieste vysoko nad oblakmi. Nemenná vec, toto svetlo. Vzdialené Slnko, ktoré toho svetlo reprezentovalo, o päť miliárd rokov dohorí; ale teraz, v *tejto* chvíli, Slnko stále svieti. A to sa nikdy nezmení. Načo si želať, aby veci zostali navždy rovnaké, keď toto želanie už bolo splnené tak absolútne, ako len želanie môže byť? Paradox stálosti a nestálosti: iba z pohľadu toho druhého existovali veci ako pokrok, alebo strata.

„Vždy si mi radila dobre,“ povedal Jeffreyssai. „Toto sa nemenilo. Celý čas, čo sa poznáme.“

Sklonila hlavu, prikývla. Toto bola pravda a nebolo treba vyslovovať jej dôsledky.

„Takže,“ povedal Jeffreyssai. „Nie kvôli hádke. Iba preto, že chcem poznať odpoveď. *Si* si istá?“ Ani nevedel, ako by to mohla *odhadnúť*.

„Dosť istá,“ povedala, „dlhú dobu sme zhromažďovali štatistiky, a v deväťsto osemdesiatich piatich prípadoch z tisíca ako je ten tvoj...“

Potom sa zasmiala na výraze jeho tváre. „Nie, žartujem. Samozrejme, že si nie som istá. Túto vec vieš rozhodnúť iba ty. Ale *som* si istá, že by si si mal dať pauzu a urobiť to, čo robievate – som si dosť istá, že na to máte nejaký rituál, aj keď o ňom nehovoríte s nezasvätenými – keď *veľmi vážne zvažujete* zanechanie dlhodobého predpokladu vašej existencie.“

Bolo naozaj ťažké hádať sa s tým, premýšľal Jeffreyssai, najmä keď vám odborník na danú oblasť povedal, že sa pravdepodobne mýlite.

→ http://lesswrong.com/lw/ur/crisis_of_faith/

„Uznávam," povedal Jeffreyssai. Z jeho úst mala táto veta konečnosť príkazu. *Nie je nutné sa do mnou ďalej hádať: Vyhrala si.*

„Ale prestaň," povedala. Vstala z vankúša jedným plynulým posunom bez najlepšieho zbytočného pohybu. Nevystatovala sa svojím vekom, ale ani ho neskrývala. Prijala jeho vystretú ruku a zdvihla ju k perám na formálny bozk. „Zbohom, sensei."

„Zbohom?" zopakoval Jeffreyssai. To naznačovalo vyšší stupeň rozchodu než *dovidenia*. „Mám v úmysle ťa opäť navštíviť, milady; a ty si tu vždy vítaná."

Prešla ku dverám bez odpovede. Vo dverách sa zastavila, bez otočenia. „Už to nebude to isté," povedala. A potom bez toho, že by jej pohyby vyzerali čo len trochu náhlivo, odišla tak rýchlo, že to bolo takmer akoby zmizla.

Jeffreyssai si vzdychol. Ale prinajmenšom odteraz po samotnú skúšku boli jeho činy predpísané, známe množstvá.

Opustil túto formálnu prijímaciu miestnosť, prešiel cez svoju arénu a poslal za svojimi žiakmi poslov s odkazom, že zajtrajšie lekcie musia improvizovať v jeho neprítomnosti, a že neskôr bude skúška.

A potom nerobil nič konkrétne. Prečítal si ďalších sto strán učebnice, ktorú mal požičanú; nebola veľmi dobrá, ale napokon kniha, ktorú požičal na výmenu tiež nebola veľmi dobrá. Túlal sa z jednej izby vo svojom dome do druhej, potulujúc sa kontroloval rôzne sklady, aby videl, či sa niečo ukradlo (chýbal balíček kariet, ale to bolo všetko). Z času na čas sa jeho myšlienky vrátili k zajtrajšej výzve, a on ich nechal plynúť. Vôbec svoje myšlienky neusmerňoval, iba zablokoval každú myšlienku, ktorá mu napadla *predtým*; a nedovolil žiaden druh záveru, ani žiadnu myšlienku ohľadom toho, kam jeho myšlienky možno smerujú.

Slnko zapadlo, a on ho chvíľu pozoroval, myseľ starostlivo nečinná. Bol to čin fantastickej rovnováhy dostať svoju myseľ do nečinnosti bez toho, že tým človek bol posadnutý, alebo že by vynakladal energiu, aby ju takto udržal; pred pár rokmi by sa z toho bol zapotil, ale cvičenie ho už dávno zdokonalilo.

Ďalšie ráno sa zobudil s chaosom z nočného sna čerstvo v mysli, a robiac, čo vedel, aby si tento pocit chaosu udržal aj v pamäti, zišiel schodiskom, potom ďalším schodiskom, potom ešte ďalším schodiskom, až nakoniec prišiel do najmenej módnej izby v celom svojom dome.

Bola biela. A to bol koniec celej jej farebnej schémy.

Pozdĺž jednej celej steny boli plakety, ktoré si podľa starej a odporúčanej metódy mladý Jeffreyssai veľmi pozorne sám narysoval, vypaľujúc si *pojmy* do svojej mysle každým dotykom štetca, ktorý písal tieto slová. *To, čo je možné zničiť pravdou, by malo byť zničené. Ľudia dokážu zniesť to, čo je pravda, pretože to už znášajú. Zvedavosť sa snaží zničiť sama seba.* Jedna malá plaketa dokonca neukazovala nič, iba červený vodorovný rez. Symboly mohli znamenať *čokoľvek*; pružnosť vizuálnej moci, pred ktorej priamym priznaním by zaváhala ešte aj Konšpirácia Bardov.

Pod plaketami boli do steny vyryté dve sady značiek. V stĺpci plus, dve značky. V stĺpci mínus, päť značiek. Sedemkrát vstúpil do tejto miestnosti; päťkrát sa rozhodol nezmeniť svoj názor; dvakrát odišiel ako iný človek. Neexistoval žiaden predpísaný pomer, ani rozsah – to by bolo výsmechom. Ale ak po čase neboli v plusovom stĺpci žiadne značky, mohli ste rovnako dobre priznať, že nemá zmysel mať túto miestnosť, pretože nemala schopnosť, ktorú predstavovala. Buď to, alebo ste sa narodili poznajúc pravdu a správnosť všetkého.

Jeffreyssai si sadol, nie čelom k plaketám, ale čelom od nich, smerom k prázdnej bielej stene. Bolo lepšie nemať vizuálne rozptýlenie.

V mysli si zopakoval meta-mnemoniku a potom rôzne odkazované sub-mnemoniky siedmych hlavných princípov a šesťdesiatich dvoch konkrétnych techník, ktoré sa s najväčšou pravdepodobnosťou mohli ukázať potrebné pri Rituále Zmeny Názoru. K týmto si Jeffreyssai pridal ďalšiu mnemoniku, ktorá mu pripomínala jeho vlastných štrnásť najzahanbujúcejších prehliadnutí.

Nenadýchol sa zhlboka. Pravidelné dýchanie bolo najlepšie.

A potom si položil tú otázku.

*→

→ http://lesswrong.com/lw/us/the_ritual/

Kniha III.

Stroj v duchovi

Mysle: Úvod

(napísal Rob Bensinger)

Vy ste myseľ, a to vás stava do pomerne zvláštneho postavenia.

Veľmi máloktoré veci sú mysle. Vy ste tá čudná časť vecí v tomto vesmíre, ktorá dokáže zostavovať predpovede a robiť plány, zvažovať a meniť názory, trpieť, snívať, sledovať lienky, alebo cítiť náhlu chuť na mango. Môžete si dokonca vytvoriť *vo vnútri svojej mysle* obrázok celej vašej mysle. Môžete rozmýšľať o svojom vlastnom procese rozmýšľania, a pracovať na tom, aby ste jeho operácie priviedli do väčšieho súladu s vašimi cieľmi.

Vy ste myseľ, uskutočnená v ľudskom mozgu. A ukazuje sa, že ľudský mozog, pri všetkej svojej úžasnej pružnosti, je vec, ktorá sa riadi zákonmi, vec vzorcov a programov. Vaša myseľ sa dokáže riadiť nejakým programom po celý svoj život, a ani raz si nevšimnúť, že to robí. A tieto programy môžu mať veľké dôsledky.

Keď vám nejaký myšlienkový vzorec dobre slúži, nazývame to „rozumnosť“.

Existujete taký, aký ste, so zabudovanými príkladmi určitých druhov rozumnosti a určitých druhov nerozumnosti, vďaka svojmu pôvodu. Vy, a všetok život na Zemi, pochádzate z pravekých molekúl, ktoré sa dokázali replikovať. Tento proces replikácie bol spočiatku ťarbavý a náhodný, a čoskoro priniesol replikovateľné *rozdiely* medzi replikátormi. „Evolúcia“ je naše označenie pre postupné zmeny v týchto rozdieloch.

Keďže niektoré z týchto reprodukovateľných rozdielov majú dopad na reprodukovateľnosť – jav označovaný ako „výber“ - evolúcia viedla k organizmom vhodným na rozmnožovanie sa v prostrediach, aké mali ich predkovia. Všetko vo vás je postavené na ozvenách zápasov a víťazstiev vašich predkov.

A tak ste tu: myseľ, vytesaná zo slabších myslí, snažiaca sa pochopiť svoje vlastné vnútorné fungovanie, aby ho mohla zdokonaliť – zdokonaliť vzhľadom na *vaše* ciele, a nie ciele vášho tvorcu, evolúcie. Aké užitočné pravidlá a vhľady si môžeme vziať z poznania, že toto je naša základná situácia?

Duchovia a stroje

Naše mozgy, vo svojej štruktúre a dynamike na malej škále, vyzerajú ako mnohé iné mechanické systémy. Napriek tomu zriedkavo rozmýšľame o svojich mysliach pomocou rovnakých pojmov ako uvažujeme o predmetoch v našom prostredí alebo orgánoch v našom tele. Naše základné myšlienkové kategórie – názor, rozhodnutie, slovo, myšlienka, pocit, a tak ďalej – sa málo podobajú na naše fyzické kategórie.

Filozofi v minulosti vzali toto pozorovanie a prijali ho, argumentujúc, že mysle a mozgy sú principiálne odlišné a oddelené javy. Toto je pohľad, ktorý filozof Gilbert Ryle nazval „dogma o Duchovi v Stroji“.[146] Ale moderní vedci a filozofi, ktorí odmietli dualizmus, ho nemuseli nutne nahradiť lepším predvídajúcim modelom fungovania mysle. *Prakticky* povedané, naše ciele a túžby stále fungujú ako voľne sa vznášajúci duchovia, ako magistérium oddelené od zvyšku nášho vedeckého poznania. Môžeme hovoriť o „rozumnosti“ a „skreslení“ a „ako zmeniť svoj názor“, ale ak sú tieto myšlienky stále nepresné

146 Gilbert Ryle, *The Concept of Mind* (University of Chicago Press, 1949).

a neobmedzené zastrešujúcou teóriou, náš vedecky znejúci jazyk nám neochráni pred robením rovnakých druhov chýb ako boli tie, ktorých teoretické východiská zahŕňali duchov a esencie.

Čo je zaujímavé, tajomstvo a mystifikácia obklopujúce mysle nezahmlievajú iba náš pohľad na *ľudí*. Prirastajú aj k systémom, ktoré sa zdajú podobné mysliam alebo účelné v evolučnej biológii a umelej inteligencii (UI). Možno, ak nedokážeme pohotovo pochopiť, čo sme, z pohľadu na seba samých, môžeme sa naučiť viac použitím očividne *nie* ľudských procesov ako zrkadla.

Môžeme sa tu učiť od mnohých duchov – duchov minulosti, súčasnosti i budúcnosti. A tieto ilúzie sú skutočné kognitívne udalosti, skutočné javy, ktoré môžeme študovať a vysvetliť. Ak sa *zdá*, že v stroji je nejaký duch, samotné toto zdanie je skrytou činnosťou daného stroja.

Prvá postupnosť v *Stroj v duchu*, „Jednoduchá matematika evolúcie“, má za cieľ komunikovať nesúlad a odchýlky medzi našou dedičnou históriou, našou súčasnou biológiou, a našimi konečnými ambíciami. Toto bude vyžadovať ponoriť sa hlbšie než je zvykom v úvodoch do evolúcie pre nebiológov, ktoré často obmedzujú svoju pozornosť na povrchné črty prirodzeného výberu.

Tretia postupnosť „Ľudská príručka k slovám“ opisuje základný vzťah medzi poznaním a tvorbou pojmov. Po tomto nasleduje dlhšia esej uvádzajúca do bayesovského odvodzovania.

Ako most medzi týmito témami „Krehké ciele“ abstrahuje z ľudského vedomia a evolúcie k myšlienke myslí a na cieľ zameraných systémov vo všeobecnosti. Tieto eseje slúžia sekundárne na vysvetlenie autorovho všeobecného prístupu k filozofii a vede rozumnosti, ktorý je silne ovplyvnený jeho prácou na UI.

Prebudovanie inteligencie

Yudkowsky je teoretik rozhodovania a matematik pracujúci na základných témach všeobecnej umelej inteligencie (VUI), teoretickom štúdiu systémov riešiacich problémy v rôznych oblastiach. Yudkowskeho práca na UI bola hlavným motorom za jeho skúmaním psychológie ľudskej rozumnosti, ako poznamenáva vo svojom celkom prvom článku na blogu *Overcoming Bias*, Bojové umenie rozumnosti:

> To poznanie rozumnosti, ktoré mám, som nadobudol v procese zápasenia s problémom všeobecnej umelej inteligencie (čo je úloha, ktorej úspešné zavŕšenie by si vyžadovalo dostatočné majstrovstvo v rozumnosti, aby ste dokázali zostrojiť hotového fungujúceho racionalistu zo špáradiel a gumičiek). Vo väčšine stránok je problém UI omnoho náročnejší než osobné umenie rozumnosti, ale v niektorých veciach je v skutočnosti ľahší. V bojovom umení mysle potrebujeme nadobudnúť procedurálne zručnosti v reálnom čase potiahnuť tie správne páky na veľkom, už existujúcom stroji na myslenie, ktorého vnútornú štruktúru nemôže konečný používateľ upravovať. Niečo z tohto stroja je optimalizované na tlaky evolučného výberu smerujúce priamo proti našim zamýšľaným cieľom pri jeho používaní. Vedome sa rozhodujeme, že chceme hľadať iba pravdu; ale naše mozgy majú zabudovanú podporu na racionalizovanie nepravdy. (…)
>
> Pokúsiť sa zostrojiť osobné umenie rozumnosti, použitím vedy o rozumnosti, sa môže ukázať ťažkopádne: Človek si predstavuje, že skúsi vymyslieť bojové umenie pomocou abstraktnej teórie fyziky, teórie hier, a ľudskej anatómie. Lenže ľudia nie sú reflexívne slepí; máme prirodzený inštinkt pre sebapozorovanie. Naše vnútorné oko nie je slepé; ale vidí rozmazane, so systematickými odchýlkami. Potrebujeme teda aplikovať vedu na naše intuície, použiť abstraktné poznanie na napravenie svojich myšlienkových pohybov, a rozšírenie svojich metakognitívnych zručností. Nepíšeme počítačový program, aby sme primäli bábku vykonávať zostavy bojového umenia; sú to naše vlastné myšlienkové končatiny, ktorými musíme hýbať. Preto musíme spojiť teóriu s praxou. Musíme uvidieť, čo tá veda znamená, pre nás, pre náš každodenný vnútorný život.

Domnievam sa, že z Yudkowskeho pohľadu je hovoriť o ľudskej rozumnosti a nepovedať pritom nič zaujímavé o UI asi rovnaké náročné ako hovoriť o UI a nepovedať pritom nič zaujímavé o rozumnosti.

V dlhodobom hľadisku Yudkowsky predpovedá, že UI prekoná ľudí pri „explózii inteligencie“, scenári, kde sebamodifikujúca UI zlepší svoju vlastnú schopnosť produktívne sa redizajnovať, čo naštartuje rýchlu postupnosť ďalších sebazdokonalení. Namiesto pojmu „explózia inteligencie“ sa občas používa aj „technologická singularita“; do januára 2013 sa MIRI nazýval „inštitút singularity pre umelú inteligenciu“ a usporiadaval každoročný summit singularity. Odvtedy však Yudkowsky začal dávať prednosť staršiemu pojmu I. J. Gooda, „explózia inteligencie“, aby pomohol odlíšiť svoje názory od iných futuristických predpovedí ako je napríklad téza Raya Kurzweila o exponenciálnom technologickom pokroku.[147]

Zdá sa, že technológie ako UI inteligentnejšia než človek pravdepodobne spôsobia veľké spoločenské otrasy, či už k lepšiemu alebo k horšiemu. Yudkowsky zaviedol pojem „teória Priateľskej UI“, ktorým označuje skúmanie techník ako zladiť preferencie VUI s preferenciami ľudí. V tomto bode sa málo vie o tom, kedy by mohol byť vynájdený všeobecne inteligentný softvér, alebo aké bezpečnostné prístupy by v takýchto prípadoch mohli dobre fungovať. Súčasné autonómne UI môže byť veľmi náročné overiť a potvrdiť s dostatočnou istotou, a mnohé súčasné techniky sa pravdepodobne nebudú dať zovšeobecniť na inteligentnejšie a prispôsobivejšie systémy. „Priateľská UI“ má preto bližšie k zverincu základných matematických a filozofických otázok než k jasne danej množine programátorských cieľov.

V roku 2015 o Yudkowskeho názoroch na budúcnosť UI naďalej debatujú technologickí prognostici a výskumníci UI v priemysle a akadémii, ale zatiaľ sa neuzniesli na konsenze. Kniha Nicka Bostroma *Superinteligencia* poskytuje hrubý súhrn mnohých morálnych a strategických otázok, ktoré nastoľuje UI inteligentnejšia než človek.[148]

Na všeobecný úvod do oblasti UI je najčastejšie používanou učebnicou *Umelá inteligencia: moderný prístup* od Russella a Norviga.[149] V kapitole rozoberajúcej morálne a filozofické otázky nastolené UI, Russell a Norvig poznamenávajú o technickej náročnosti špecifikovania dobrého správania u silne prispôsobivej UI:

> [Yudkowsky] tvrdí, že priateľskosť (túžba neublížiť ľuďom) by mala byť nadizajnovaná už od začiatku, ale že by si dizajnéri mali zároveň uvedomovať, že ich vlastné dizajny môžu byť chybné, aj že robot sa bude časom učiť a vyvíjať. Preto je to problém dizajnu mechanizmu – definovať mechanizmus pre vyvíjanie systémov UI pod systémom kontrol a poistiek, a dať týmto systémom funkcie úžitku, ktoré zostanú priateľské aj pri takýchto zmenách. Nemôžeme dať programu jednoducho statickú funkciu úžitku, pretože okolnosti, aj naše želané reakcie na okolnosti, sa časom menia.

Znepokojení možnosťou, že budúci pokrok v UI, nanotechnológii, biotechnológii, a ďalších oblastiach môže ohroziť ľudskú civilizáciu, Bostrom a Ćirković zostavili prvú akademickú antológiu na danú tému, *Riziká globálnych katastrof*.[150] Najextrémnejšie z nich sú *existenčné riziká*, riziká, ktoré môžu spôsobiť trvalú stagnáciu alebo vyhynutie ľudstva.[151]

Ľudia (vrátane odborníkov) majú sklon byť *mimoriadne slabí* v predpovedaní veľkých udalostí v budúcnosti (vrátane nových technológií). Časť Yudkowskeho cieľa pri diskusii o rozumnosti je zistiť, ktoré skreslenia prekážajú našej schopnosti predpovedať veľké otrasy a dobre sa na ne pripraviť. Yudkowskeho príspevky k zbierke *Riziká globálnych katastrof* „Kognitívne skreslenia potenciálne

147 Irving John Good, „Speculations Concerning the First Ultraintelligent Machine,“ in *Advances in Computers*, ed. Franz L. Alt and Morris Rubinoff, vol. 6 (New York: Academic Press, 1965), 31–88, doi:10.1016/S0065-2458(08)60418-0.

148 Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014).

149 Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Upper Saddle River, NJ: Prentice-Hall, 2010).

150 Bostrom and Ćirković, *Global Catastrophic Risks*.

151 Príkladom možného existenčného rizika je scenár „šedej kaše“, v ktorom molekulárni roboti navrhnutí tak, aby sa vedeli efektívne kopírovať, urobia svoju prácu až príliš dobre, a rýchlo vytlačia živé organizmy, ako skonzumujú všetku dostupnú hmotu na Zemi.

ovplyvňujúce hodnotenie globálnych rizík“ a „Umelá inteligencia ako pozitívny a negatívny faktor globálneho rizika“ spájajú jeho výskum v kognitívnej vede a v UI. Yudkowsky a Bostrom zhŕňajú krátkodobé otázky s dlhodobými v kapitole *Cambridgskej príručky o umelej inteligencii* „Etika umelej inteligencie“.[152]

Hoci je toto kniha o *ľudskej* rozumnosti, téma UI je relevantná ako zdroj jednoduchých ilustrácií stránok ľudského myslenia. Dlhodobé technologické predpovede sú tiež jedným z dôležitejších uplatnení bayesovskej rozumnosti, ktorá dokáže modelovať správne uvažovanie aj v oblastiach, kde je údajov málo alebo sú rozporné.

Poznať dizajn vám môže povedať veľa o dizajnérovi; a poznať dizajnéra vám môže povedať veľa o dizajne.

Začnime teda skúmaním, čo nás náš vlastný dizajnér dokáže naučiť o nás samotných.

Medzihra: Sila inteligencie

Vo svojich lebkách nosíme asi kilo slizkého, vlhkého, sivého tkaniva, zvlneného ako pokrčený toaletný papier. Keby ste pozreli na túto nechutnú hrču, nepomysleli by ste si, že je to jedna z najmocnejších vecí v známom vesmíre. Keby ste nikdy nevideli učebnicu anatómie, a videli by ste mozog ležať na ulici, povedali by ste „Fuj!“ a pokúsili by ste sa nezamazať si od neho topánky. Aristoteles si myslel, že mozog je orgán na chladenie krvi. *Vyzerá* neškodne.

Pred piatimi miliónmi rokov predkovia levov vládli cez deň a predkovia vlkov sa túlali nocou. Vládnuce šelmy boli vyzbrojené zubami a pazúrmi – ostrými, tvrdými čepeľami, ovládanými mocnými svalmi. Ich korisť si v sebaobrane vyvíjala pancierové schránky, ostré rohy, jedy, maskovanie. Táto vojna trvala tisíce vekov a nespočetne veľa pretekov v zbrojení. Nejeden porazený vypadol z hry, ale nikde nebolo stopy po víťazovi. Kde jeden druh mal pancier, iný sa vyvinul, aby ho prelomil; kde sa jeden druh stal jedovatým, iný si vyvinul odolnosť voči jedu. Každý druh mal svoje vlastné miesto – veď kto by dokázal žiť v mori, na oblohe, a na zemi zároveň? Neexistovala dokonalá zbraň ani dokonalá obrana, ani dôvod veriť, že čosi také je možné.

Potom prišiel čas Mäkkých Vecí.

Nemali brnenie. Nemali pazúre. Nemali jed.

Keby ste videli film zobrazujúci jadrový výbuch, a keby vám povedali, že to urobil jeden zo živočíchov na Zemi, ani v najdivokejšom sne by ste si nepredstavovali, že za to môžu Mäkké Veci. Napokon, Mäkké Veci nie sú rádioaktívne.

Na začiatku Mäkké Veci nemali ani stíhačky, ani guľomety, ani pušky, ani meče. Ani bronz, ani železo. Ani kladivá, ani nákovy, ani kliešte, ani kováčske dielne, ani bane. Mäkké Veci mali akurát mäkké prsty – príliš slabé na to, aby zlomili strom, tobôž horu. Očividne neškodné. Na sekanie kameňa by bolo treba oceľ, ale Mäkké Veci nevedeli vylučovať oceľ. V prírode neboli žiadne oceľové čepele, ktoré by ich mäkké prsty mohli zdvihnúť. Ich telá nedokázali vytvoriť ani zďaleka dostatočnú teplotu na roztavenie kovu. Celý scenár bol jasne absurdný.

A čo sa týka Mäkkých Vecí manipulujúcich DNA – to už by ani nebolo smiešne. Mäkké prsty nie sú také malé. Na ich úrovni nie je DNA dostupná; to by bolo ako pokúšať sa zdvihnúť atóm vodíka. Iste, *technicky vzaté* všetko tvorí jeden vesmír, *technicky vzaté* sú Mäkké Veci a DNA časti toho istého sveta, s jednotnými zákonmi fyziky, rovnakou veľkou sieťou kauzality. Ale buďme realistickí: odtiaľto tam sa dostať nedá.

Aj keby si Mäkké Veci *jedného dňa* dokázali vyvinúť niektorú z týchto schopností, trvalo by im to milióny rokov. Celé veky sme sledovali príliv a odliv Života, a môžeme vám povedať, jeden rok, to nie je ani len tiknutie na hodinách evolučného času. Isteže, *technicky vzaté*, jeden rok je šesťsto biliónov biliónov biliónov Planckových intervalov. Ale za menej ako šesťsto miliónov biliónov biliónov biliónov biliónov Planckových intervalov sa aj tak nikdy nič nestane, takže táto námietka je iba

152 Nick Bostrom and Eliezer Yudkowsky, „The Ethics of Artificial Intelligence,“ in *The Cambridge Handbook of Artificial Intelligence*, ed. Keith Frankish and William Ramsey (New York: Cambridge University Press, 2014).

akademická. Mäkké Veci, ktoré práve bežia krížom savanou, nebudú lietať krížom cez kontinent prinajmenšom ďalších desať miliónov rokov; *nikto* nemôže mať tak veľa sexu.

A teraz mi opäť vysvetlite, prečo umelá inteligencia nemôže urobiť nič zaujímavé cez internet, dokiaľ jej človek programátor nepostaví robotické telo.

Všimol som si, že automatická reakcia človeka na slovo „inteligencia“ - to, čo mu vyvstane v mysli v prvej polovici sekundy po počutí slova „inteligencia“ - často určuje jeho automatickú reakciu na singularitu. Často sa im pod slovom „inteligencia“ vynára pojem *nepraktického intelektuála* – predstava šachového veľmajstra, ktorý si nevie nájsť priateľku, alebo univerzitného profesora, ktorý by mimo akademickej pôdy nedokázal prežiť.

„Aby ste uspeli v práci, potrebujete viac než len inteligenciu,“ hovoria ľudia, ako keby sila osobnosti sídlila v obličkách, a nie v mozgu. „Inteligencia je bezmocná voči puške,“ hovoria, ako keby pušky rástli na stromoch. „Kde by umelá inteligencia získala peniaze?“ pýtajú sa, ako keby prvému *Homo sapiens* spadli dolárové bankovky z neba a on ich mohol použiť v supermarkete už existujúcom v pralese. Ľudský druh sa nenarodil do trhovej ekonomiky. Včely vám nepredajú med, ak im zaň ponúknete platbu kreditkou. Ľudský druh peniaze *vymyslel*, a peniaze existujú – pre nás, nie pre myši alebo osy – pretože v ne naďalej veríme.

Stále sa pokúšam vysvetliť ľuďom, že vzorom inteligencie nie je Dustin Hoffman vo filme *Rain Man*, ale že je to človek ako taký, bodka. Je to mäkká vec, ktorá vybuchne vo vákuu, avšak zanechala odtlačky nôh na svojom mesiaci. Vnútri tej sivej hrče je schopnosť prehľadávať cesty cez veľkú sieť kauzality, a nájsť trasu vedúcu k napohľad nemožnému – schopnosť, ktorú niekedy nazývame tvorivosť.

Ľudia – špeciálne investori – sa občas pýtajú, ako, ak Inštitút výskumu strojovej inteligencie úspešne zostaví skutočnú umelú inteligenciu, bude možné výsledky *speňažiť*. To je to, čo nazývame problém uhla pohľadu.

Alebo to je možno niečo hlbšie než obyčajná zrážka predpokladov. S trochou tvorivého myslenia si ľudia dokážu predstaviť, ako by sa dalo cestovať na Mesiac, alebo vyliečiť kiahne, alebo vyrábať počítače. Predstaviť si však trik, ktorý by dokázal dosiahnuť *všetky tieto veci zároveň*, sa im zdá celkom nemožné – napriek tomu, že takáto sila sídli iba pár centimetrov za ich vlastnými očami. Tá sivá vlhká hmota sa stále zdá byť záhadnou samotnej tejto sivej vlhkej hmote.

A tak, keďže ľudia nedokážu celkom vidieť, ako by to celé mohlo fungovať, sa sila inteligencie zdá menej skutočná. Predstaviť si ju je ťažšie než predstaviť si vežu z ohňa, ktorá pošle loď na Mars. Plán návštevy Marsu upútava našu predstavivosť. Ale keby niekto sľuboval aj návštevu Marsu, aj veľkú zjednotenú teóriu fyziky, aj dôkaz Riemannovej hypotézy, aj liek proti obezite, aj liek proti rakovine, aj liek proti starnutiu, aj liek proti hlúposti... to skrátka znie nesprávne.

A veru by aj malo. Je vážnym zlyhaním predstavivosti myslieť si, že inteligencia dokáže iba tak málo. Kto by si dokázal predstaviť, vtedy dávno, čo budú jedného dňa mysle robiť. Možno ešte ani *nevieme*, čo sú naše skutočné problémy.

Ale medzičasom, keďže je ťažké vidieť, ako môže mať jeden proces takéto rozmanité schopnosti, je ťažké si predstaviť, že by sa zároveň vyriešili čo len také bežné problémy ako obezita a rakovina a starnutie.

Lenže, jeden trik už raz vyliečil kiahne a postavil lietadlá a kultivoval pšenicu a skrotil oheň. Naša súčasná veda možno nemá jasno v tom, ako presne ten trik funguje, ale on napriek tomu funguje. Ak dočasne nerozumieme nejakému javu, je to fakt o danom stave našej mysle, nie fakt o danom jave. Prázdna mapa neznamená, že aj územie je prázdne. Aj keď človek celkom nerozumie tej sile, ktorá zanechala odtlačky nôh na Mesiaci, tie odtlačky tam napriek tomu stále sú – skutočné odtlačky, na skutočnom Mesiaci, umiestnené tam skutočnou silou. Keby človek tejto sile dostatočne hlboko porozumel, mohol by ju vytvárať a usmerňovať. Inteligencia je rovnako skutočná ako elektrina. Akurát je omnoho mocnejšia, omnoho nebezpečnejšia, má omnoho hlbšie dopady na pokračovanie rozvíjajúceho sa príbehu života vo vesmíre... a je o trochu ťažšie dôjsť na to, ako postaviť jej generátor.

* →

→ http://yudkowsky.net/singularity/power/

L: Jednoduchá matematika evolúcie

131. Cudzí boh

„Evolučná teória má jednu zvláštnu vlastnosť," povedal Jacques Monod, „každý si myslí, že jej rozumie."

Keď sa človek pozerá na svet prírody, tisíckrát vidí *účel*. Zajačie nohy, vytvorené a prispôsobené na beh; líščie čeľuste, vytvorené a prispôsobené na trhanie. Ale to, čo vidíte, nie je presne to, čo tam je...

V časoch pred Darwinom bola príčina všetkej tejto zdanlivej *účelnosti* pre vedu veľmi veľkou záhadou. Zástancovia Boha hovorili: „urobil to Boh", pretože vždy, keď ste vo vete použili slovo „Boh", dostali ste 50 kladných bodov. Ale možno teraz nie som férový. V časoch pred Darwinom to vyzeralo ako omnoho rozumnejšia hypotéza. Ak nájdete v púšti hodinky, povedal William Paley, dokážete odvodiť existenciu hodinára.

Keď sa však pozriete na *všetku* túto napohľad účelnosť v prírode, namiesto vyberania si príkladov, ktoré sa vám hodia, začnete si všímať veci, ktoré nezapadajú do židovsko-kresťanského konceptu dobrotivého Boha. Líšky vyzerajú dobre navrhnuté, aby chytali zajace. Zajace vyzerajú dobre navrhnuté, aby unikali líškam. Vari mal Stvoriteľ problém rozhodnúť sa, čo vlastne chce?

Keď navrhujem hriankovač, nenavrhujem jednu časť, ktorá sa pokúša dostať elektrinu do cievky a druhú časť, ktorá sa pokúša zabrániť elektrine dostať sa do cievky. To by bolo plytvanie úsilím. *Kto* navrhol ekosystém s jeho lovcami a korisťami, vírusmi a baktériami? Ešte aj kaktus, o ktorom si možno myslíte, že je dobre navrhnutý, aby poskytoval vodnatú dužinu púštnym zvieratám, je pokrytý nepohodlnými pichliačmi.

Ekosystém by dával omnoho viac zmyslu, keby ho nevytvoril jeden *Niekto*, ale skôr horda božstiev – povedzme hinduistických alebo šintoistických. To by ľahko vysvetlilo aj všadeprítomnú účelnosť aj všadeprítomný konflikt: konalo tu viac božstiev, občas s rozdielnym zámerom. Aj líška aj zajac sú navrhnuté, ale rôznymi súperiacimi božstvami. Ktovie, či si niekto všimol túto napohľad skvelú indíciu v prospech hinduizmu oproti kresťanstvu. Pravdepodobne nie.

Podobne, židovsko-kresťanský Boh je údajne láskavý – no, svojím spôsobom. A predsa mnohé z *účelov* prírody vyzerajú priam kruto. Darwin mal podozrenie, že Stvoriteľ nie je štandardný, keď študoval osu Ichneumon, ktorá uštipnutím paralyzuje svoju obeť, a tú potom zaživa zožerú jej larvy: „Nedokážem uveriť," napísal Darwin, „že by láskavý a všemocný Boh mohol navrhnúť Ichneumonidae s výslovným úmyslom, aby sa kŕmili žijúcimi telami húseníc, alebo mačky, aby sa hrali s myšami." Ktovie, či si niektorí dávnejší myslitelia všimli túto skvelú indíciu v prospech manichejských náboženstiev oproti monoteistickým.

Dnes už všetci poznáme pointu: Skrátka poviete: „evolúcia".

Obávam sa, že takto niektorí ľudia prijímajú „vedecké" vysvetlenie, ako čarovnú továreň prírody na účelnosť. Už som spomínal prípad Storm z filmu *X-Men*, ktorá v *jednej mutácii* získa schopnosť vrhať blesky. Prečo? No, lebo existuje taká vec, ktorá sa volá „evolúcia" a tá nejako do prírody napumpuje veľa účelnosti, a tieto zmeny sa dejú cez „mutácie". Takže ak Storm dostane naozaj *veľkú* mutáciu, môže sa zmeniť tak, že hádže blesky. Super obľúbeným zdrojom je rádioaktivita: žiarenie spôsobuje mutácie, takže silnejšie žiarenie spôsobí silnejšie mutácie. To je logické.

Lenže evolúcia nedovoľuje do prírody preniknúť len tak *hocijakému* druhu účelnosti. To je to, čo robí evolúciu úspešnou empirickou hypotézou. Keby evolúcia dokázala vysvetliť aj hriankovač, nielen strom, bola by bezcenná. Evolučná teória je viac než len ukazovanie prstom na prírodu a hovorenie: „Teraz je dovolený účel" alebo „Urobila to evolúcia!" Sila teórie nie je v tom, čo dovoľuje, ale v tom, čo zakazuje; ak dokážete vymyslieť rovnako presvedčivé vysvetlenie pre každý výsledok, máte nulové vedomosti.

„Mnohí nebiológovia," všimol si George Williams, „si myslia, že to pre ich dobro rastú štrkáčom na konci chvosta štrkadlá." Bzzz! *Takýto* druh účelnosti nie je dovolený. Evolúcia nefunguje tak, že sa náhodne objavujú záblesky účelnosti – jeden živočíšny druh sa zmení v prospech náhodného adresáta.

Evolúciu poháňa systematická korelácia medzi rôznymi spôsobmi ako rôzne gény vytvárajú organizmy a koľko kópií týchto génov sa dostane do ďalšej generácie. Aby štrkáčom narástli štrkadlá, musia byť gény na rast štrkadiel stále častejšie v každej nasledujúcej generácii. (V skutočnosti sú to gény na čoraz zložitejšie štrkadlá, ale ak začnem opisovať všetky finesy a zádrhely evolučnej biológie, naozaj tu *budeme* celý deň.)

Neexistuje víla Evolúcia, ktorá sa pozrie na súčasný stav prírody, rozhodne sa, čo by bol „dobrý nápad“ a vyberie si, že zvýši frekvenciu génov pre stavbu štrkadiel.

Obávam sa, že na tomto mieste sa veľa ľudí v evolučnej biológii zasekne. Rozumejú, že „užitočné“ gény sa stávajú častejšími, ale slovo „užitočné“ im dovolí vsunúť akýkoľvek účel. Nemyslia si, že by existovala víla Evolúcia, ale predsa sa pýtajú, ktoré gény sú „užitočné“, akoby gény štrkáčov mohli byť „užitočné“ nie-štrkáčom.

Základ je uvedomiť si, že víla Evolúcia neexistuje. Neexistuje vonkajšia sila, ktorý by *rozhodovala*, ktoré gény treba propagovať. Všetko, čo sa stane, sa stane *kvôli* týmto génom samotným.

Gény na stavbu (čoraz lepších) štrkadiel sa museli nejako stať stále častejšími v genofonde štrkáčov *kvôli* štrkadlám. V tomto prípade *pravdepodobne* preto, lebo štrkáče s lepšími štrkadlami častejšie prežili – namiesto aby sa úspešnejšie rozmnožovali, alebo mali bratov, ktorí sa úspešnejšie rozmnožujú, atď.

Možno si dravce dávajú pozor na štrkadlá a tak nestúpia na hada. Alebo možno štrkadlo odvádza pozornosť od hadovej hlavy. (Ako naznačuje George Williams: „Výsledok boja medzi psom a zmijou výrazne závisí na tom, či pes pôvodne schmatol zmiju za hlavu alebo za chvost.“)

Ale to je len hadie štrkadlo. Existujú aj zložitejšie spôsoby, ako môže gén spôsobiť, že sa jeho kópie stanú v nasledujúcej generácii častejšími. Váš brat alebo sestra má s vami spoločnú polovicu génov. Gén, ktorý obetuje jednu jednotku prostriedkov, aby tak poskytol tri jednotky prostriedkov bratovi, môže pomáhať niektorým svojim kópiám tým, že obetuje jeden z jeho zostrojených organizmov. (Ak naozaj chcete vedieť všetky finesy a zádrhely, kúpte si knihu o evolučnej biológii; neexistuje tu kráľovská cesta.)

Pointa je, že to *účinok génu* musí *spôsobiť*, že sa kópie tohto génu stanú častejšími v nasledujúcej generácii. Neexistuje žiadna víla Evolúcia, ktorá by zasahovala zvonka. Neexistuje nič, čo sa *rozhoduje*, že niektoré gény sú „užitočné“ a preto by sa ich frekvencia mala zvýšiť. Je to iba príčina a následok, začínajúc od samotných génov.

Toto vysvetľuje čudnú konfliktnú účelnosť v prírode a jej častú krutosť. Vysvetľuje to ešte lepšie než horda šintoistických božstiev.

Prečo je toľko častí prírody vo vojne s inými časťami prírody? Pretože neexistuje jedna Evolúcia, ktorá by riadila celý tento proces. Existuje toľko rôznych „evolúcií“, koľko je reprodukujúcich sa populácií. Zajačie gény sa stávajú častejšími alebo zriedkavejšími v zajačích populáciách. Líščie gény sa stávajú častejšími alebo zriedkavejšími v líščích populáciách. Líščie gény vytvárajúce líšky, ktoré chytajú zajace, umiestnia viac svojich kópií do nasledujúcej generácie. Zajačie gény vytvárajúce zajace, ktoré ujdú líškam, budú *prirodzene* častejšie v nasledujúcej generácii zajacov. Preto slovné spojenie: „prirodzený výber“.

Prečo je príroda krutá? Vy ako človek sa môžete pozrieť na osu Ichneumon a rozhodnúť sa, že je kruté jesť svoju obeť zaživa. Môžete sa rozhodnúť, že ak už idete jesť svoju obeť zaživa, budete mať aspoň toľko slušnosti, aby ste ju zbavili bolesti. Sotva by to osu niečo stálo, aby svoju obeť zároveň s paralyzovaním aj anestetizovala. Alebo čo staré slony, ktoré zomierajú od hladu, keď im vypadnú posledné zuby? Tieto slony sa aj tak už nebudú rozmnožovať. Čo by to evolúciu stálo – presnejšie, evolúciu slonov – zabezpečiť, že ten slon zomrie hneď, namiesto pomaly a bolestivo? Čo by to evolúciu stálo dať tomu slonovi pred smrťou anestetikum alebo príjemné sny? Nič; ten slon by sa tak či tak nerozmnožoval o nič viac ani menej.

Keby ste sa rozprávali s iným človekom, pokúšali sa vyriešiť konflikt záujmov, boli by ste v dobrej vyjednávacej pozícii – ľahko by ste ho presvedčili. Stálo by to tak málo anestetizovať tú korisť, nechať slona zomrieť bez agónie! Ach, prosím, mohli by ste to láskavo urobiť... ehm....

Nemáte však s kým vyjednávať.

Ľudia si vymýšľajú falošné zdôvodnenia; pomocou jednej metódy zistia, čo chcú, a potom si to zdôvodnia pomocou inej metódy. Neexistuje žiadna víla Slonia Evolúcia, ktorá by (a) zisťovala, čo je pre slony najlepšie, a potom (b) hľadala, ako to *zdôvodniť* pred Evolučným Dozorcom, ktorý (c) nechce, aby sa schopnosť rozmnožovať znížila, ale (d) je ochotný pristúpiť na myšlienku bezbolestnej smrti, pokiaľ to nebude škodiť žiadnym génom.

Nikde v tomto systéme nie je žiaden obhajca slonov.

Ľudia, ktorých často hlboko znepokojuje blahobyt zvierat, môžu veľmi presvedčivo argumentovať, prečo rôzne láskavosti vôbec nebudú škodiť schopnosti rozmnožovať sa. Žiaľ, evolúcia slonov nepoužíva podobný algoritmus; nevyberá pekné gény, *o ktorých sa dá presvedčivo argumentovať*, že pomôžu schopnosti rozmnožovať sa. Jednoducho: gény, ktoré sa častejšie rozmnožujú sa stanú častejšími v nasledujúcej generácii. Ako keď voda tečia dole kopcom, rovnako láskavo.

Človek pri pohľade na prírodu začne uvažovať o tom, ako by sme *my* navrhli organizmy. A potom má sklon začať si racionalizovať, prečo by jeho vylepšenia dizajnu zvýšili schopnosť rozmnožovať sa – je to politický inštinkt, pokus predať svoju obľúbenú možnosť ako to, čo je v súlade so šéfovým obľúbeným zdôvodnením.

A tak amatérski evoluční biológovia vymýšľajú rôzne druhy úžasných a *celkom pomýlených* predpovedí. Pretože tí amatérski biológovia nakreslili svoje spodné riadky – a čo je dôležitejšie, našli svoje predpovede v priestore hypotéz – pomocou *iného* algoritmu než používa príroda na nakreslenie *svojho* spodného riadku.

Človek inžinier by navrhol ľudské chuťové bunky tak, aby merali, koľko máme ktorých živín a koľko ich potrebujeme. Keby bolo málo tuku, mandle alebo cheeseburgery by chutili lahodne. Ale keby ste začali byť obézni, alebo keby vám chýbali vitamíny, šalát by chutil lahodne. Lenže neexistuje víla Ľudská Evolúcia, ktorý by inteligentne plánovala a navrhla by všeobecný systém na každý prípad. V pravekom prostredí bolo spoľahlivo nemenné, že kalórie boli vzácne. Preto sa gény, ktorých organizmy milovali kalórie, stali častejšími. Ako keď voda tečie dole kopcom.

My sme jednoducho stelesnením histórie toho, ktoré organizmy *naozaj* prežili a rozmnožili sa, nie ktoré organizmy *by po obozretnej úvahe mali* prežiť a rozmnožiť sa.

Ľudská sietnica je postavená naopak: Svetlocitlivé bunky sú vzadu a nervy vychádzajú vpredu a potom idú naspäť *cez* sietnicu do mozgu. Preto slepá škrvna. Človeku inžinierovi toto pripadá jednoducho hlúpe – a iným organizmom sa sietnice nezávisle vyvinuli správnym spôsobom. Prečo nezmeniť návrh sietnice?

Problém je, že žiadna *jednotlivá* mutácia neotočí naraz *celú* sietnicu. Človek inžinier môže zároveň zmeniť návrh viacerých častí alebo v pláne zohľadniť budúce zmeny. Ale ak jednotlivá mutácia pokazí nejakú dôležitú časť organizmu, nezáleží na tom, aké úžasne veci by na nej víla dokázala postaviť – organizmus zomrie a frekvencia génov sa zníži.

Ak otočíte bunky sietnice bez preprogramovania nervov a optických káblov, systém ako celok nebude fungovať. Nezáleží na tom, že pre vílu alebo človeka inžiniera by to bol krok vpred pri predizajnovaní sietnice. Ten organizmus je slepý. Evolúcia nemá predvídavosť; je to jednoducho zmrazená história toho, ktoré organizmy sa *naozaj* rozmnožili. Evolúcia je rovnako slepá ako tá napoly predizajnovaná sietnica.

Ak nájdete v púšti hodinky, povedal William Paley, dokážete odvodiť existenciu hodinára. Boli raz takí, ktorí to popierali, ktorí si mysleli, že život „len tak vznikal“ bez nejakého optimalizačného procesu, že myši spontánne vznikali zo slamy a špinavých tričiek.

Ak sa opýtame, kto bol *bližšie* k pravde – či teológovia, ktorí argumentovali Bohom Stvoriteľom, alebo intelektuálne nenaplnení ateisti, ktorí argumentovali, že myši vznikajú spontánne – potom musíme za víťazov vyhlásiť teológov: evolúcia nie je Boh, ale má bližšie k Bohu než k čistej náhodnej entropii. Mutácia je náhodná, ale výber je nenáhodný. To neznamená, že tam zasahuje a vyberá inteligentná víla. Znamená to, že je nenulová štatistická korelácia medzi génom a tým, ako často sa organizmus rozmnožuje. Počas miliónov rokov sa tieto nenulové štatistické korelácie skladajú do niečoho veľmi mocného. Nie je to boh, ale viac sa to podobá na boha než na sneženie na televíznej obrazovke.

V mnohom *je* evolúcia v súlade s teológiou. „Bohovia sú ontologicky odlišní od stvorení," povedal Damien Broderick, „inak nie sú hodní toho papiera, na ktorom sú napísaní." A naozaj, sám Formovač Života nie je stvorenie. Evolúcia je nehmotná, tak ako židovsko-kresťanské božstvo. Všadeprítomná v prírode, skrytá v páde každého listu. Rozľahlá ako povrch planéty. Miliardy rokov stará. Sama nestvorená, prirodzene vyplývajúca zo štruktúry fyziky. Neznie azda toto všetko ako niečo, čo by sa mohlo povedať o Bohu?

A predsa tento Tvorca nemá žiaden rozum, rovnako ako nemá telo. V niektorých ohľadoch má jeho práca neuveriteľne slabý dizajn podľa ľudských merítok. Je vnútorne rozdelený. A najmä, nie je *dobrý*.

V istom zmysle Darwin *objavil* Boha – Boha, ktorý nespĺňal predsudky teológie a tak prišiel bez zvestovania. Keby Darwin objavil, že život stvoril inteligentný činiteľ – nehmotná myseľ, ktorá nás miluje a zabije nás bleskom, ak sa opovážime povedať opak – ľudia by povedali: „Páni! To je Boh!"

Namiesto toho však Darwin objavil čudného cudzieho Boha – nie pohodlne „nevýslovného", ale *naozaj skutočne odlišného od nás*. Evolúcia nie je Boh, ale keby bola, nebol by to Jehova. Bol by to Azathoth od H. P. Lovecrafta, slepý šialený Boh chaoticky bublajúci v strede všetkého, obkolesený tenkým monotónnym pískaním fláut.

Čo ste mohli predpovedať, keby ste sa naozaj *pozreli* na prírodu.

Toľko teda k tvrdeniu niektorých religionistov, že veria v nejasné božstvo s primerane vysokou pravdepodobnosťou. Každý, kto by *naozaj* veril v nejasné božstvo, by bol rozoznal svojho čudného neľudského stvoriteľa, keď Darwin povedal: „Aha!"

Toľko teda k tvrdeniu niektorých religionistov, že s nevinnou zvedavosťou čakajú, až veda objaví Boha. Veda už objavila v istom zmysle božského stvoriteľa ľudí – ale to nebolo to, čo chceli religionisti počuť. Čakali na objavenie *svojho* Boha, veľmi konkrétneho Boha, ktorého tam *oni* chceli mať. Budú teda čakať naveky, pretože ten veľký objav *už nastal*, a víťazom je Azathoth.

Nuž, tým viac moci pre nás, ľudí. Páči sa mi mať Stvoriteľa, ktorého viem prekabátiť. Lepšie než byť domácim zvieratkom. Som rád, že to bol Azathoth a nie Odin.

* →

132. Zázrak evolúcie

Zázrakom evolúcie je to, že vôbec funguje.

Myslím to doslova: Ak chcete žasnúť nad evolúciou, tak presne nad týmto sa oplatí žasnúť.

Ako vznikne prvá optimalizácia vo vesmíre? Ak inteligentný činiteľ navrhol prírodu, kto navrhol tohto inteligentného činiteľa? Kde je prvý dizajn, ktorý nemá dizajnéra? Záhada *nie je* v tom, ako môže byť prvá etapa tohto procesu super chytrá a super efektívna; záhada je, ako sa to *vôbec* môže stať.

Evolúcia rieši problém nekonečnej regresie nie tým, že by bola super chytrá a super efektívna, ale tým, že je hlúpa a neefektívna a napriek tomu funguje. *Toto* je ten zázrak.

Z pracovných dôvodov občas musím diskutovať o pomalosti, náhodnosti a slepote evolúcie. Potom niekto povie: „*Ty* si práve povedal, že evolúcia nedokáže naplánovať viaceré zmeny zároveň, a že evolúcia je veľmi neefektívna, lebo mutácie sú náhodné. Nie je toto to, čo hovoria *kreacionisti*? Že nemôžeš zložiť hodinky tým, že náhodne zatrasieš súčiastkami v krabici?"

Lenže odpoveď kreacionistom nie je to, že *môžete* zložiť hodinky tým, že zatrasiete súčiastkami v krabici. Odpoveď je, že evolúcia funguje *inak*. Ak si myslíte, že evolúcia funguje pomocou tornád, ktoré skladajú lietadlá 747, potom sa kreacionistom podarilo nesprávne vám vysvetliť biológiu; predali vám slameného panáka.

Skutočná odpoveď je, že zložité mechanizmy sa vyvíjajú buď postupne, alebo prispôsobovaním predchádzajúcich zložitých mechanizmov novým účelom. Veverice skáču zo stromu na strom pomocou svalov, ale dĺžka ich skoku do istej miery závisí od aerodynamiky ich tiel. Dnes už teda existujú lietajúce veverice, také aerodynamické, že dokážu preplachtiť krátku vzdialenosť. Keby vtáky vyhynuli,

→ http://lesswrong.com/lw/kr/an_alien_god/

potomkovia lietajúcich veveríc by za desať miliónov rokov mohli znovu osídliť ich ekologickú niku, premeniac plachtiace membrány na krídla. A kreacionisti by hovorili: „Načo je dobrá polovica krídla? Iba spadneš a rozpleštíš sa. Ako sa vevericovtáky mohli vyvinúť postupne?“

Takto jedna zložitá adaptácia môže pomôcť naštartovať novú zložitú adaptáciu. Zložitosť sa tiež môže hromadiť postupne, začínajúc od jednej mutácie.

Najprv príde nejaký gén A, ktorý je jednoduchý, ale aspoň *trochu* užitočný, takže A sa postupne rozšíri do celého genofondu. Teraz príde gén B, ktorý je užitočný iba v prítomnosti génu A, ale keďže A je v genofonde spoľahlivo prítomný, prirodzený výber tlačí v prospech B. Teraz vznikne modifikovaná verzia A*, ktorá závisí na B, ale nenaruší závislosť B na A/A*. Potom príde C, ktoré závisí na A* a B, a B*, ktoré závisí na A* a C. O chvíľu máte „neredukovateľne zložitý“ mechanizmus, ktorý sa pokazí, keď z neho odoberiete ľubovoľnú časť.

Napriek tomu si stále môžete predstaviť cestu späť k tomu jedinému kúsku: môžete, bez pokazenia celého stroja urobiť jeden kúsok menej závislým na druhom kúsku, a urobiť toto niekoľkokrát, až nakoniec môžete jeden kúsok celý vybrať *bez* pokazenia stroja, a tak ďalej, až kým z tikajúcich hodiniek neurobíte primitívne slnečné hodiny.

Tu je príklad: DNA ukladá informácie veľmi pekne v trvanlivom formáte, ktorý umožňuje presné kopírovanie. Ribozóm premieňa túto uloženú informáciu na postupnosť aminokyselín, bielkovinu, ktorá sa poskladá na rôzne druhy chemicky aktívnych tvarov. Tento spojený systém, DNA a ribozóm, dokáže zostaviť rôzne druhy bielkovinových strojov. Ale načo je dobrá DNA bez ribozómu, ktorý premieňa informáciu v DNA na bielkoviny? Načo je dobrý ribozóm bez DNA, ktorá mu povie, aké bielkoviny má vyrobiť?

Organizmy nezanechávajú vždy skameneliny a evoluční biológovia nedokážu *vždy* vypátrať cestu postupných zlepšení. Ale v tomto prípade *vieme*, ako sa to stalo. RNA má s DNA spoločnú tú vlastnosť, že dokáže prenášať informácie a kopírovať sa, hoci RNA je menej trvanlivá a kopíruje sa menej presne. A RNA má s bielkovinami spoločnú tú vlastnosť, že sa dokáže skladať do chemicky aktívnych tvarov, hoci nie je taká pružná ako bielkovinové reťazce aminokyselín. Takmer isto je RNA tým spoločným A, ktoré predchádzalo navzájom závislým A* a B.

Dôležité je všimnúť si nielen to, že RNA robí spoločnú prácu DNA a bielkovín *horšie*, ale aj to, že túto spoločnú prácu vôbec robí. Je dosť úžasné, že *jedna molekula* dokáže zároveň ukladať informácie aj manipulovať chémiu. Aby to celé robila *dobre*, to už je celkom nepotrebný zázrak.

Čo bolo prvým replikátorom na svete? Mohlo to byť vlákno RNA, pretože *istou zhodou okolností* sa na Zemi pred vznikom života, pred 4 miliardami rokov, chemické zložky RNA vyskytovali. Prosím všimnite si: evolúcia *nevysvetľuje* vznik života; nie je *úlohou* evolučnej biológie vysvetliť prvý replikátor, pretože *prvý* replikátor nepochádza z iného replikátora. Evolúcia popisuje štatistické trendy pri replikovaní. Prvý replikátor nebol štatistický trend, bola to číra náhoda. Predstava, že evolúcia musí vysvetliť *pôvod* života je čistý slamený panák – ďalšie kreacionistické zavádzajúce tvrdenie.

Keby ste boli sledovali prvotnú polievku v deň prvého replikátora, v deň, ktorý zmenil celú Zem, neurobilo by na vás dojem *ako dobre* sa prvý replikátor replikuje. Prvý replikátor sa pravdepodobne replikoval ako opitá opica na LSD. Neprejavoval by žiadne príznaky starostlivého vyladenia prítomné v moderných replikátoroch, pretože prvý replikátor bol *náhoda*. Nebolo *potrebné*, aby sa toto jedno vlákno RNA alebo chemický hypercyklus alebo vzor v hline kopíroval elegantne. Stačilo, že sa to *vôbec* dialo. Ešte aj tak to asi bolo veľmi nepravdepodobné, ako izolovaná udalosť – ale stačilo, aby sa to stalo *raz* a prílivových jazierok bolo veľa. O pár miliárd rokov neskôr sa replikátory prechádzajú po mesiaci.

Prvý náhodný replikátor bol tá najdôležitejšia molekula všetkých čias. Ale keby ste ju príliš vychvaľovali, pripisovali jej všemožné úžasné schopnosti pomáhajúce replikácii, *unikla by vám celá pointa*.

Nemyslite si, že v politickom boji medzi evolucionistami a kreacionistami ten, kto vychvaľuje evolúciu, musí byť na strane vedy. Veda má veľmi presnú predstavu o schopnostiach evolúcie. Ak chválite evolúciu čo len o milimeter viac, nie je to „boj na strane evolúcie“ proti kreacionizmu. Je to vedecky nepresné, bodka. Padnete do kreacionistickej pasce, ak začnete tvrdiť, že áno, tornádo *má*

schopnosť poskladať lietadlo 747! Aké úžasné! Aká úžasne inteligentná je evolúcia, aká úctyhodná! Pozrite na mňa, ako prisahám vernosť vede! Čím viac pekných vecí poviem o evolúcii, tým viac musím byť na strane evolúcie proti kreacionistom!

Lenže prílišné chválenie evolúcie ničí ten *skutočný* zázrak, ktorým nie je to, *ako dobre* evolúcia navrhuje veci, ale to, že prirodzene prebiehajúci proces *vôbec* dokáže niečo navrhovať.

Zbavme sa teda predstavy, že evolúcia je úžasný dizajnér alebo úžasný sprievodca druhov na ceste za ich osudom, ktorého by ľudia mali napodobňovať. Keby ľudská inteligencia napodobňovala evolúciu ako dizajnéra, to by bolo ako keby sa sofistikovaná moderná baktéria pokúšala napodobňovať prvý replikátor ako biochemika. Ako povedal „Darwinov buldog" T. H. Huxley:[153]

> Pochopme už raz a navždy, že etický pokrok spoločnosti nespočíva v napodobňovaní vesmírneho procesu a už vôbec nie v utekaní pred ním, ale v boji proti nemu.

Huxley to nepovedal preto, že by neveril v evolúciu, ale preto, lebo jej veľmi dobre rozumel.

*→

133. Evolúcie sú hlúpe (a predsa fungujú)

V predchádzajúcej kapitole som napísal:

> Veda má veľmi presnú predstavu o schopnostiach evolúcie. Ak chválite evolúciu čo len o milimeter viac, nie je to „boj na strane evolúcie" proti kreacionizmu. Je to vedecky nepresné, bodka.

V tejto kapitole opíšem niektoré všeobecne známe neefektívnosti a obmedzenia evolúcií. Používam „evolúcie" v množnom čísle, pretože evolúcia líšok funguje v konflikte s evolúciou zajacov, a žiadna z nich sa nedokáže porozprávať s evolúciou hadov, aby sa od nej naučila stavať jedovaté tesáky.

Hovorím tu teda o obmedzeniach evolúcie, ale to neznamená, že sa sem snažím prepašovať kreacionizmus. Toto je štandardná evolučná biológia pre druhákov. (Pre piatakov, ak musíte odvodiť dané rovnice.) Takto ohraničené evolúcie stále dokážu vysvetliť pozorovanú biológiu; tie obmedzenia sú v skutočnosti nutné, aby to dávalo zmysel. Pamätajte, že zázrak evolúcií nie je v tom, ako dobre fungujú, ale že vôbec fungujú.

Ľudská inteligencia je taká zložitá, že nikto nemá dobrý spôsob ako odmerať, aká je efektívna. Prirodzený výber, hoci nie je jednoduchý, je *jednoduchší než* ľudský mozog; a primerane pomalší a menej efektívny, ako je primerané pre prvý optimalizačný proces, ktorý kedy existoval. Evolúcie sú v skutočnosti dosť jednoduché na to, aby sme vedeli spočítať, *ako presne sú hlúpe*.

Evolúcie sú pomalé. Aké pomalé? Predstavme si, že existuje užitočná mutácia, ktorý poskytuje výhodu spôsobilosti 3 %: v priemere majú nositelia tohto génu 1,03-krát viac detí ako nenositelia. Za predpokladu, že sa táto mutácia vôbec rozšíri, ako dlho jej potrvá, než sa rozšíri do celej populácie? To závisí od veľkosti populácie. Gén, ktorý poskytuje 3% výhodu prispôsobenia, šíriaci sa v populácii 100 000 jedincov, potrebuje v priemere 768 generácií, aby sa stal v genofonde všadeprítomným. Populácia 500 000 jedincov potrebuje 875 generácií. Všeobecný vzorec je:

$$\text{Generácie do zafixovania} = 2 \ln(N) / s$$

kde N je veľkosť populácie a (1 + s) je spôsobilosť. (Ak má každý nositeľ génu 1,03-krát viac detí ako nenositeľ, s = 0,03.)

Ak by teda populácia mala veľkosť 1 000 000 – odhadovaná populácia v časoch lovcov-zberačov – trvalo by 2763 generácií, aby sa gén poskytujúci 1% výhodu rozšíril do celého genofondu.[154]

To by nemalo byť prekvapujúce; gény si celé svoje rozšírenie musia sami odpracovať. Neexistuje žiadna víla Evolúcia, ktorá by sa pozrela na genofond a povedala: „Hm, zdá sa, že tento gén sa rýchlo šíri

153 Thomas Henry Huxley, *Evolution and Ethics and Other Essays* (Macmillan, 1894).

→ http://lesswrong.com/lw/ks/the_wonder_of_evolution/

154 Dan Graur and Wen-Hsiung Li, *Fundamentals of Molecular Evolution*, [Základy molekulárnej evolúcie] 2nd ed. (Sunderland, MA: Sinauer Associates, 2000).

– mala by som ho nadeliť každému.“ V ľudskej trhovej ekonomike, ak niekto legitímne získava 20% výnos z investície – najmä ak je za tým jasný a jednoduchý mechanizmus – môže rýchlo získať ďalší kapitál od ostatných investorov; a ďalší začnú zakladať rovnaké podniky. Gény sa musia šíriť bez pomoci búrz alebo bánk alebo napodobňovateľov – ako keby Henry Ford musel vyrobiť jedno auto, predať ho, nakúpiť súčiastky na 1,01 auta (v priemere), predať tieto autá, a pokračovať v tom, až kým nedosiahne milión áut.

Toto všetko predpokladá, že sa gén vôbec rozšíri. Táto rovnica je jednoduchšia a nezávisí od veľkosti populácie:

Pravdepodobnosť zafixovania = 2 s

Mutácia, ktorá dáva výhodu 3 % (čo je na pomery mutácií sakra veľa) má šancu 6 %, že sa rozšíri, prinajmenšom v tomto konkrétnom prípade.[155] Mutácie môžu nastať viac než raz, ale v miliónovej populácii s presnosťou kopírovania 10^{-8} chýb za bázu za generáciu môžete čakať sto generácií na ďalšiu príležitosť, ktorá opäť bude mať iba šancu 6 %, že sa zafixuje.

Napriek tomu, *z dlhodobého hľadiska* má evolúcia dobrú šancu, že sa tam nakoniec dostane. (Táto téma sa tu bude opakovať.)

Zložitým adaptáciám trvá *veľmi* dlho, kým sa vyvinú. Najprv príde alela A, ktorá je výhodná aj osamote, a vyžaduje tisíc generácií, aby sa zafixovala v genofonde. *Až* potom sa môže iná alela B, ktorá závisí od A, pokúsiť o zafixovanie. Hrubý kožuch nie je silnou výhodou, pokiaľ vaše životné prostredie nemá *štatisticky spoľahlivý* sklon púšťať na vás studené počasie. Nuž, súčasťou prostredia pre gény sú aj iné gény, a ak B závisí od A, potom B nebude mať silnú výhodu, dokiaľ sa A *spoľahlivo* nevyskytuje v genetickom prostredí.

Povedzme, že B poskytuje výhodu 5 % v prítomnosti A, inak žiadnu výhodu. Potom, dokiaľ má A ešte v populácii frekvenciu iba 1 %, B poskytuje svoju výhodu v 1 prípade zo 100, čiže priemerná výhoda prispôsobenia z B je 0,05% a pravdepodobnosť, že sa B zafixuje, je 0,1 %. Pri zložitej adaptácii sa *najprv* musí A vyvíjať tisíc generácií, *potom* sa musí B vyvíjať ďalších tisíc generácií, *potom* sa A* vyvíja ďalších tisíc generácií... a o niekoľko miliónov rokov neskôr máme zložitú adaptáciu.

A ostatné evolúcie ju nenapodobňujú. Ak evolúcia hadov vyvinula úžasný nový jed, nepomôže to evolúcii líšok ani evolúcii levov.

Porovnajme si toto s človekom programátorom, ktorý dokáže vytvoriť nový zložitý mechanizmus so stovkami navzájom závisiacich častí za *jediné popoludnie*. Ako je to vôbec *možné*? Nepoznám celú odpoveď a odhadujem, že ju nepozná ani veda; ľudské mozgy sú omnoho zložitejšie než evolúcie. Mohol by som zamávať rukami a povedať niečo ako: „na cieľ zamerané spätné reťazenie využívajúce kombinačné modulárne reprezentácie“, ale to by vám nepomohlo navrhnúť si svojho vlastného človeka. Ale aj tak: Ľudia dokážu predvídavo navrhovať nové časti v očakávaní neskoršieho návrhu ďalších nových častí; produkovať koordinované zmeny vo viacerých nezávislých mechanizmoch zároveň; učiť sa pozorovaním jednotlivých pokusných prípadov; zamerať sa na problémové miesta a myslieť abstraktne na ich riešenie; a určovať si priority, ktoré zmeny sa oplatí skúšať, namiesto čakania až dopadajúci vesmírny lúč vytvorí niečo dobré. Z pohľadu prirodzeného výberu je toto jednoducho mágia.

Ľudia dokážu robiť veci, ktoré by evolúcie pravdepodobne nedokázali urobiť, *bodka*, ani za celú odhadovanú životnosť vesmíru. Ako raz povedala známa biologička Cynthia Kenyon pri večeri, na ktorej som mal česť sa zúčastniť: „Jeden doktorand dokáže za hodinu urobiť veci, ktoré evolúcia nedokáže urobiť za miliardu rokov.“ Podľa súčasných najlepších vedomostí biológov evolúcie vynašli úplne rotujúce koleso celkovo *trikrát*.

A nezabudnite na tú časť, kde programátor uverejní časti svojho kódu na internete.

Áno, niektoré diela evolúcie sú pôsobivé aj v porovnaní s najlepšou technológiou *Homo sapiens*. Lenže naša kambrijská explózia iba začala, skutočne sme začali akumulovať vedomosti iba približne...

155 John B. S. Haldane, „A Mathematical Theory of Natural and Artificial Selection,“ [Matematická teória prirodzeného a umelého výberu] *Mathematical Proceedings of the Cambridge Philosophical Society* 23 (5 1927): 607–615, doi:10.1017/S0305004100011750.

pred koľkými, štyristo rokmi? V niektorých veciach biológia stále prevyšuje najlepšie ľudské technológie: nedokážeme postaviť samoreplikujúci systém veľkosti motýľa. V iných veciach ľudská technológia zanecháva biológiu pozadu. Máme kolesá, máme oceľ, máme pušky, máme nože, máme oštepy, máme rakety, máme tranzistory, máme jadrové elektrárne. Každým ďalším desaťročím sa rovnováha nakláňa ďalej.

Takže, ešte raz: aby sa človek pozeral na prirodzený výber ako na inšpiráciu v umení dizajnu, to je ako keby sa sofistikovaná moderná baktéria pokúšala napodobniť biochémiu prvého trápneho replikátora. Keby sa prvý replikátor objavil v dnešnej konkurenčnej ekológii, bol by okamžite zožraný. Rovnaký osud by čakal človeka plánovača, ktorý by skúšal robiť vo svojej stratégii náhodné zmeny a čakať 768 iterácií testovania, aby prijal zlepšenie o 3 %.

Nechváľte evolúcie ani o milimeter viac než si zaslúžia.

Očakávajte čoskoro: Ďalšie vzrušujúce matematické obmedzenia evolúcie!

*→

134. Korporácie ani nanoprístroje nemajú evolúciu

> Fyzikálne zákony ani matematické pravidlá neprestanú pôsobiť. To ma vedie k názoru, že evolúcia nezastane. To ma ďalej vedie k názoru, že príroda – s krvavými tesákmi a pazúrmi, ako ju niektorí označili – sa jednoducho dostane na ďalšiu úroveň...
>
> [Pokúsiť sa zbaviť Darwinovskej evolúcie je] ako pokúsiť sa zbaviť gravitácie. Dokiaľ existujú obmedzené zdroje a viacerí súperiaci činitelia schopní odovzdávať svoje vlastnosti, máme selekčný tlak.
>
> --Perry Metzger, predpovedajúci, že vláda prirodzeného výberu bude pokračovať do nekonečna

V evolučnej biológii, tak ako v mnohých iných oblastiach, je dôležité myslieť viac kvantitatívne než kvalitatívne. Šíri sa užitočná mutácia „občas, ale nie vždy“? Nuž, zázračná schopnosť by bola užitočnou mutáciou, takže by ste čakali, že sa bude šíriť, nie? Lenže toto je kvalitatívne uvažovanie, nie kvantitatívne – ak je X pravda, potom je Y pravda; ak sú zázračné schopnosti užitočné, možno sa budú šíriť. V kapitole Evolúcie sú hlúpe som opísal rovnice pravdepodobnosti fixácie užitočnej mutácie, zhruba dvojnásobok výhody v spôsobilosti (6 % pre výhodu 3 %). Iba tento druh číselného myslenia nám pravdepodobne pomôže uvedomiť si, že mutácie, ktoré sú *užitočné iba zriedkavo*, majú veľmi malú šancu sa rozšíriť, a že je prakticky nemožné, aby vznikla zložitá adaptácia bez ustavičného používania. Keby zázračné schopnosti naozaj existovali, museli by sme očakávať, že ich všetci budú každú chvíľu používať – nie preto, že by to bolo úžasne užitočné, ale preto, lebo bez toho by sa v prvom rade nemohli vyvinúť.

„Dokiaľ existujú obmedzené zdroje a viacerí súperiaci činitelia schopní odovzdávať svoje vlastnosti, máme selekčný tlak.“ Toto je kvalitatívne uvažovanie. *Koľko* selekčného tlaku?

Existuje niekoľko kandidátov na najdôležitejšiu rovnicu evolučnej biológie, ale ja by som vybral Priceovu rovnicu, ktorá vo svojej najjednoduchšej podobe znie:

zmena priemernej vlastnosti = kovariancia(relatívna spôsobilosť, vlastnosť)

Toto je *veľmi* silný a všeobecný vzorec. Napríklad ak ako Z, vlastnosť, ktorá sa mení, označíme konkrétny gén pre výšku, potom Priceov vzorec hovorí, že zmena pravdepodobnosti vlastnenia daného génu sa rovná kovariancii daného génu s reprodukčnou spôsobilosťou. Alebo môžeme ako Z vziať *výšku ako takú*, bez ohľadu na konkrétne gény, a Priceov vzorec hovorí, že zmena výšky v nasledujúcej generácii sa rovná kovariancii výšky s relatívnou reprodukčnou spôsobilosťou.

(Prinajmenšom to platí, ak je výška priamočiaro dedičná. Ak sa zlepší výživa, takže rovnaký genotyp sa stane vyšším, musíte do Priceovej rovnice pridať opravný člen. Ak je medzi mnohými génmi

→ http://lesswrong.com/lw/kt/evolutions_are_stupid_but_work_anyway/

zložitý nelineárny vzťah, buď musíte pridať opravný člen alebo počítať túto rovnicu takým zložitým spôsobom, že prestane byť objasňujúca.)

Mnoho osvietenia možno dosiahnuť štúdiom rôznych foriem a odvodení Priceovej rovnice. Napríklad záverečná rovnica hovorí, že priemerná vlastnosť sa mení podľa svojej *kovariancie* s *relatívnou* spôsobilosťou, a nie s *absolútnou* spôsobilosťou. To znamená, že ak gén Froda zachráni celý druh pred vyhynutím, priemerná vlastnosť Froda sa nezvýši, pretože Frodov čin bol rovnako užitočný všetkým genotypom a teda *nekovarioval* s *relatívnou* spôsobilosťou.

Hovorí sa, že Pricea dôsledky jeho rovnice pre altruizmus natoľko rozrušili, že spáchal samovraždu, ale možno mal aj iné problémy. (*Overcoming Bias* neodporúča páchať samovraždu po štúdiu Priceovej rovnice.)

Jedno z osvietení, ktoré môžete získať po meditácii nad Priceovou rovnicou je, že „obmedzené zdroje“ a „viacerí súperiaci činitelia schopní odovzdávať vlastnosti“ *nestačia* na vznik evolúcie. „Veci, ktoré sa replikujú“ nie je dostatočná podmienka. Dokonca ani „súťaž medzi replikujúcimi sa vecami“ nestačí.

Vyvíjajú sa korporácie? Určite spolu súperia. Občas sa z nich oddelia potomkovia. Ich prostriedky sú obmedzené. Občas zanikajú.

Lenže nakoľko sa potomok korporácie podobá na svojich rodičov? Veľká časť osobnosti korporácie sa odvíja od jej kľúčových funkcionárov, a riaditelia sa nerozdeľujú štiepením. Priceova rovnica funguje iba natoľko, ako sa vlastnosti dedia v generáciách. Pokiaľ sa pra-pra-prapotomkovia nepodobajú na svojich pra-pra-prarodičov, nedostanete viac selekčného tlaku než za štyri generácie – čokoľvek, čo sa udialo pred viac ako štyrmi generáciami, bude vymazané. Áno, osobnosť korporácie môže ovplyvnuť jej osamostatnenú pobočku – ale to nie je ako dedičnosť DNA, ktorá je digitálna namiesto analógovej, a dokáže sa prenášať s 10^{-8} chybami za bázu za generáciu.

V DNA vám dedičnosť vydrží *milióny* generácií. Vďaka tomu môžu púhou evolúciou vzniknúť zložité adaptácie – digitálna DNA vydrží dosť dlho na to, aby sa gén dávajúci výhodu 3 % rozšíril počas 768 generácií, a potom môže vzniknúť nový gén, ktorý na ňom závisí. Dokonca aj keby sa korporácie rozmnožovali s digitálnou vernosťou, zatiaľ by boli nanajvýš desiatou generáciou vo svete RNA.

Iste, korporácie sú *selektované*, v tom zmysle, že neschopné korporácie bankrotujú. To by logicky malo spôsobiť, že budeme s väčšou pravdepodobnosťou pozorovať korporácie s vlastnosťami pomáhajúcimi schopnosti. V rovnakom zmysle, keď sa z hviezdy stane nova krátko po jej vzniku, je menej pravdepodobné, že ju uvidíte na nočnej oblohe. Lenže ak náhodná hviezdna dynamika spôsobí, že jedna hviezda horí dlhšie než iný, to neznamená, že aj v budúcnosti budú hviezdy pravdepodobne horieť dlhšie – táto vlastnosť sa neskopíruje do iných hviezd. Nemali by sme očakávať, že budúci astrofyzici objavia *zložité* vnútorné ústroje hviezd, zdanlivo navrhnuté, aby im pomohli horieť dlhšie. Taký druh mechanickej adaptácie si vyžaduje omnoho väčší *kumulatívny* výber než jednorazové vytriedenie.

Spomeňte si na princíp opísaný pri Einsteinovej drzosti – že drvivá väčšina indície potrebnej na rozmýšľanie o všeobecnej relativite slúžila na to, aby sa jedna konkrétna rovnica dostala na úroveň Einsteinovej osobnej pozornosti; množstvo indície potrebné na to, aby sa z vedome zvažovanej možnosti stala 99,9% istota, bolo v porovnaní s tým triviálne. V tom istom zmysle, zložité vlastnosti korporácií, ktoré si vyžadujú stovky bitov špecifikácie, sú vytvorené v prvom rade ľudskou inteligenciou, nie hŕstkou generácií evolúcie s nízkou vernosťou. V biológii sú mutácie čisto náhodné a evolúcia pokytuje tisíce bitov kumulatívneho selekčného tlaku. Pri korporáciách ľudia ponúkajú tisíce bitov inteligentne navrhnutých zložitých „mutácií“ a nasledujúci selekčný tlak „Zbankrotovalo to alebo nie?“ zodpovedá za hŕstku dodatočných bitov vysvetľujúcich, čo vidíte.

Vyspelá molekulárna nanotechnológia – umelá, nie biologická – by sa mala dokázať kopírovať s digitálnou presnosťou tisíce generácií. Mala by tu príležitosť Priceova rovnica?

Korelácia je kovariancia delená rozptylom, čiže ak A výrazne predpovedá B, môže byť medzi nimi silná „korelácia“ dokonca aj keď je A v rozsahu od 0 do 9 a B je iba v rozsahu od 50,0001 do 50,0009. Priceova rovnica používa *kovarianciu* vlastností s reprodukciou – nie koreláciu! Ak dokážete natlačiť rozptyl vlastnosti do úzkeho pásma, kovariancia klesne, a s ňou aj kumulatívna zmena danej vlastnosti.

Foresight Institute odporúča, medzi inými rozumnými návrhmi, aby príkazy na rozmnožovanie každého nanoprístroja boli šifrované. Navyše, šifrované tak, že zmena jediného bitu zakódovanej inštrukcie úplne rozhasí výsledok dešifrovania. Ak všetky vytvorené nanoprístroje budú na molekulu presné kópie, a navyše žiadne chyby počas montáže nebudú dedičné, pretože potomkovia dostanú digitálnu kópiu pôvodných zašifrovaných príkazov na vytvorenie vnukov, potom sa vaše nanoprístroje nebudú veľmi vyvíjať.

Stále sa budete musieť báť priónov – samoreplikujúcich sa chýb montáže mimo zašifrovaných príkazov, kde rameno robota nezachytí atóm uhlíka, ktorý malo použiť pri konštrukcii svojho homológu, a to spôsobí, že rameno potomka takisto nezachytí atóm uhlíka, atď., aj keby všetky zašifrované príkazy zostali rovnaké. Ale aká bude asi *korelácia* medzi takýmto druhom prenosnej chyby a *vyššou* mierou rozmnožovania? Predpokladajme, že jeden nanoprístroj vyprodukuje svoju kópiu každých 1000 sekúnd, a že nový nanoprístroj je zázračne efektívnejší (nielenže má prión, ale má *užitočný* prión) a kopíruje sa každých 999,99999 sekúnd. Ako vidíte, stačí mu o jeden atóm uhlíka menej. To nie je zvlášť veľa rozptylu v rozmnožovaní, takže to nie je ani zvlášť veľa kovariancie.

A ako často sa tieto nanoprístroje budú potrebovať rozmnožovať? Pokiaľ nebudú mať k dispozícii viac atómov než existuje v slnečnej sústave, alebo povedzme vo viditeľnom vesmíre, prejde iba pár generácií než narazia na nedostatok zdrojov. „Obmedzené zdroje“ nie sú dostatočnou podmienkou evolúcie; potrebujete často iterovanú smrť prevažnej časti populácie, ktorá uvoľní zdroje. „Generácie“ vlastne nie sú ani tak celé čísla, ako skôr integrál nad zlomkom populácie, ktorý sa skladá z novo vytvorených jedincov.

Pre mňa je toto tá najstrašidelnejšia vec na sivom slize alebo nanotechnologických zbraniach – že by mohli pohltiť celú Zem a to by bolo *všetko*, nič zaujímavé by sa už potom nestalo. Diamant je stabilnejší než bielkoviny držané pohromade van der Waalsovými silami, takže tento sliz by si iba potreboval opraviť pár kúskov, keby dopadol asteroid. Dokonca aj keby boli prióny dostatočne silným výrazovým prostriedkom, aby sa mohli vyvíjať – evolúcia je dosť pomalá už s digitálnou DNA! - medzi pohltením Zeme slizom a vyhasnutím Slnka by uplynulo menej než 1,0 generácie.

Aby som to zhrnul, ak máte *všetky* nasledujúce vlastnosti:

- Veci, ktoré sa replikujú
- Podstatný rozptyl v ich vlastnostiach
- Podstatný rozptyl v ich rozmnožovaní
- Trvalú koreláciu medzi vlastnosťami a rozmnožovaním
- Dlhodobú dedičnosť vlastností s vysokou vernosťou
- Časté rozmnožovanie významnej časti rodičovskej populácie
- A toto *všetko* platí počas *mnohých* iterácií

Potom budete mať významné *kumulatívne* selekčné tlaky, dostatočné na vytvorenie zložitých adaptácií silou evolúcie.

*→

135. Vyvinutie k vyhynutiu

Je *veľmi* častý omyl, že evolúcia pracuje pre dobro svojho druhu. Pamätáte sa, že ste počuli niekoho hovoriť o tom, ako dva zajace splodia osem zajacov a tým „prispeli k prežitiu svojho druhu“? Moderný evolučný biológ by niečo také nikdy nepovedal; to by sa skôr rozmnožoval so zajacom.

Je to opäť jeden prípad, kde musíte myslieť na viaceré abstraktné pojmy zároveň a udržiavať ich oddelené. Evolúcia *nepracuje* s konkrétnymi jednotlivcami; jednotlivci majú také gény, s akými sa narodili. Evolúcia pracuje s reprodukujúcou sa populáciou, živočíšnym druhom, v čase. Máme prirodzený sklon myslieť si, že keď víla Evolúcia *pracuje s* druhom, musí *optimalizovať pre* tento druh. Lenže to, čo sa naozaj mení, sú frekvencie génov, a frekvencie sa nezvyšujú ani neznižujú podľa toho, nakoľko daný

→ http://lesswrong.com/lw/l6/no_evolutions_for_corporations_or_nanodevices/

gén pomáha živočíšnemu druhu ako celku. Ako neskôr uvidíme, je celkom možné, aby sa živočíšny druh vyvíjal k vyhynutiu.

Prečo sa chlapci a dievčatá rodia zhruba v rovnakom počte? (Vynechajme bláznivé krajiny, ktoré používajú umelé technológie na výber pohlavia.) Aby ste videli, prečo je to prekvapujúce, uvedomte si, že 1 samec môže oplodniť 2, 10 alebo 100 samíc; nezdá sa, že by ste potrebovali rovnaký počet samcov a samíc na zaručenie prežitia druhu. Je to ešte prekvapujúcejšie v prevažnej väčšine živočíšnych druhov, kde samec veľmi málo prispieva k výchove detí – ľudia sú v úrovni rodičovskej starostlivosti mimoriadni, ešte aj medzi primátmi. Vyvážené pomery pohlaví sa nachádzajú ešte aj u živočíšnych druhov, kde samec oplodní samicu a zmizne v hmle.

Vezmite si dve skupiny na rôznych stranách hory; v skupine A každá matka porodí 2 samcov a 2 samice; v skupine B každá matka porodí 3 samice a 1 samca. Skupina A a B budú mať rovnaký počet detí, ale skupina B bude mať o 50 % viac vnúčat a o 125 % viac pravnúčat. Mohli by ste si myslieť, že toto je významná evolučná výhoda.

Uvážte však: Čím zriedkavejšími sa samci stanú, tým reprodukčne hodnotnejšími sa stanú – nie pre *skupinu*, ale pre *jednotlivého* rodiča. Každé dieťa má ako rodičov jedného samca a jednu samicu. V každej generácii je teda celkový genetický príspevok všetkých samcov rovný celkovému genetickému príspevku všetkých samíc. Čím menej je samcov, tým väčší je individuálny genetický príspevok jedného samca. Keby všetky samice okolo vás robili to, čo je dobré pre skupinu, čo je dobré pre živočíšny druh, a rodili by 1 samca na 10 samíc, vy dokážete zhrabnúť genetický *jackpot* porodením samých samcov, z ktorých od každého budete mať (v priemere) desaťkrát viac vnúčat než od ich samičích sesterníc.

Takže hoci by skupinový výber mal uprednostniť viac dievčat, individuálny výber uprednostňuje rovnaké investície do samčích aj samičích potomkov. Keď sa pozriete na štatistiky v pôrodnici, uvidíte na prvý pohľad, že kvantitatívna rovnováha medzi silami skupinového výberu a individuálneho výberu je u *Homo sapiens* prevažne naklonená v prospech individuálneho výberu.

(Technicky povedané, nie je to jasné na prvý pohľad. Individuálny výber uprednostňuje rovnaké *rodičovské investície* do samčích a samičích potomkov. Keby porodiť a vychovať samca stálo iba polovicu, v evolučne stabilnej rovnováhe by sa rodilo dvakrát viac samcov ako samíc. Keby sa v populácii rodilo rovnaké množstvo samcov a samíc, ale porodiť samca by bolo dvakrát lacnejšie, mohli by ste opäť zhrabnúť genetický jackpot porodením väčšieho množstva samcov. Pôrodnica by teda mala odrážať rovnováhu nákladov rodičovstva, v spoločnosti lovcov a zberačov, vo výchove chlapcov a výchove dievčat; a to by ste museli nejako odhadnúť. Ale viete, nezdá sa, že by reprodukčné náklady na výchovu dievčaťa v rodine lovcov a zberačov boli o toľko *väčšie*, takže je to trochu podozrivé, že sa rodí približne rovnako veľa chlapcov a dievčat.)

Prirodzený výber nie je na skupinách, ani druhoch, dokonca ani *jednotlivcoch*. V sexuálnych druhoch sa jednotlivý organizmus nevyvíja; udržiava si rovnaké gény, s akými sa narodil. Jednotlivec je jednorazová zbierka génov, ktorá sa nikdy nezopakuje; ako chcete vyberať toto? Keď si uvedomíte, že takmer všetci vaši predkovia sú mŕtvi, je jasné, že „prežitie najspôsobilejšieho“ je mimoriadne nevhodný názov. „Replikovanie najspôsobilejšieho“ by bolo presnejšie, hoci z technického hľadiska je spôsobilosť *definovaná* iba pomocou replikácie.

Prirodzený výber je naozaj na *frekvenciách génov*. Aby sme dostali zložitú adaptáciu, stroj s mnohými nezávislými časťami, každý nový gén, ako sa vyvíja, závisí na spoľahlivej prítomnosti ostatných génov vo svojom genetickom prostredí. Musia mať vysoké frekvencie. Čím zložitejší je stroj, tým vyššie musia byť tie frekvencie. Znakom prítomnosti prirodzeného výberu je rozšírenie génu z 0,00001 % genofondu na 99 % genofondu. Toto je informácia, v zmysle teórie informácií; a toto sa musí stať, aby sa vyvinula veľká zložitá adaptácia.

Skutočným zápasom v prirodzenom výbere nie je súperenie medzi organizmami o prostriedky; to je dočasná vec, ktorej všetci účastníci zmiznú v ďalšej generácii. Tým skutočným zápasom je súperenie medzi alelami o frekvenciu v genofonde. Toto je trvalý dôsledok, ktorý vytvára trvalú informáciu. Dva barany, ktoré do seba búchajú rohmi, sú iba prchavé tiene.

Je v pohode možné, že sa alela rozšíri a zafixuje tým, že porazí alternatívnu alelu, ktorá bola „lepšia pre tento druh“. Keby Lietajúce Špagetové Monštrum magicky stvorilo druh, ktorého zmes pohlaví by bola dokonale optimalizovaná, aby zabezpečila prežitie *druhu* – optimálna zmes pohlaví, aby sa spoľahlivo vrátili zo situácii takmer vyhynutia, prispôsobili sa novým nikám, atď. - potom by evolúcia rýchlo zvrátila toto druhové optimum späť na optimum pre individuálny výber rovnakej rodičovskej investície do samcov a samíc.

Predstavte si „gén Froda“, ktorý obetuje svojho nositeľa, aby zachránil *celý jeho druh* pred vyhynutím. Čo sa v dôsledku toho stane s frekvenciou danej alely? Klesne. Díky, čau.

Keby sa hrozby vyhynutia druhu vyskytovali pravidelne (nazvime to „prostredie Buffy“), potom by sa frekvencia génu Froda systematicky znižovala, až by zanikol, a čoskoro na to aj daný druh. Hypotetický príklad? Možno. Keby mal ľudský druh zostať biologickým ešte ďalšie storočie, bol by dobrý nápad začať klonovať Gándhího.

Pri vírusoch je napätie medzi individuálnymi vírusmi rozmnožujúcimi sa tak rýchlo, ako len môžu, verzus úžitok z ponechania hostiteľa nažive dosť dlho, aby preniesol chorobu. Toto je dobrý príklad skupinového výberu v reálnom svete, a ak sa vírus vyvinie do bodu na ploche spôsobilosti, kde tlak skupinového výberu nedokáže prekonať tlak individuálneho, vírus môže krátko potom vymiznúť. Neviem, či niekto niekedy prichytil chorobu v akte vyvinutia sa k vyhynutiu, ale pravdepodobne sa to už stalo mnohokrát.

Narúšače rozdelenia podkopávajú mechanizmus, ktorý zvyčajne zaručuje férovosť pohlavného rozmnožovania. Napríklad na samčom chromozóme niektorých myší existuje narúšač rozdelenia, ktorý spôsobuje, že sa rodia iba samčie deti, všetky nesúce tento narúšač rozdelenia. Tieto samce potom oplodnia samice, ktoré porodia iba samčie deti, atď. Môžete kričať: „To je podvod!“ ale to je ľudský pohľad; reprodukčná spôsobilosť tejto alely je extrémne vysoká, keďže produkuje dvakrát viac svojich kópií v nasledujúcej generácii než jej nezmutovaná alternatíva. Dokonca aj keď sa samice stávajú vzácnejšími a vzácnejšími, samce nesúce tento gén nemajú o nič menšiu pravdepodobnosť rozmnožiť sa než hocijaké iné samce, takže narúšač rozdelenia zostáva dvakrát spôsobilejší než jeho alternatívna alela. Špekuluje sa, že skupinová selekcia v skutočnom svete mohla zohrať rolu v udržaní frekvencie tohto génu takej nízkej, ako vyzerá byť. V takom prípade, keby si myši vyvinuli schopnosť lietať a migrovať na zimu, pravdepodobne by vytvorili jednu reproduktívnu populáciu a vyvinuli by sa k vyhynutiu, ako by sa narúšač rozdelenia vyvinul k fixácii.

Okolo 50 % celkového genómu kukurice tvoria transpozóny, prvky DNA, ktorých prvoradá funkcia je kopírovať sa *na iné miesta v DNA*. Trieda transpozónov nazývaná „prvky P“ sa u drozofíl prvýkrát vyskytla asi uprostred 20. storočia a rozšírila sa do každej populácie tohto druhu za 50 rokov. „Sekvencia Alu“ u ľudí, transpozón s 300 bázami, sa opakuje medzi 300 000 a miliónkrát v ľudskom genóme. Toto nemusí druh vyhubiť, ale nepomáha mu to; transpozóny spôsobujú viac mutácií, ktoré sú takmer vždy škodlivé, znižujú efektívnu vernosť kopírovania DNA. Napriek tomu sú takíto podvodníci mimoriadne spôsobilí.

Predstavte si, že v nejakom pohlavne sa rozmnožujúcom druhu sa objavi *dokonalý* mechanizmus kopírovania DNA. Keďže väčšina mutácií je škodlivá, tento komplex génov je pre svojich nositeľov výhodný. Možno uvažujete o užitočných mutáciách – aj tie sa občas vyskytnú, neboli by teda nemutujúci v nevýhode? Lenže v pohlavnom druhu sa užitočná mutácia, ktorá vznikla u mutujúceho, môže rovnako rozšíriť aj k potomkom nemutujúcich. Mutujúci v každej generácii trpia degenerujúcimi mutáciami; a nemutujúci dokážu sexuálne nadobudnúť všetky užitočné mutácie, ktoré sa vyskytli u mutujúcich, a mať z nich tiež úžitok. Čiže mutujúci majú čistú nevýhodu. Frekvencia dokonalého mechanizmu kopírovania DNA sa rozšíri až k fixácii. O desaťtisíc rokov neskôr nastane doba ľadová a daný druh skončil. Vyvinul sa k vyhynutiu.

„Efekt okolostojaceho“ je, že keď má niekto problém, s väčšou pravdepodobnosťou zasiahne osamelý jednotlivec než skupina. Študentovi vysokej školy, ktorý predstieral epileptický záchvat, prišiel niekto na pomoc v 85 % prípadov, ak bol nablízko iba jeden svedok, a v 31 % prípadov, ak bolo nablízku päť svedkov. Uvažujem, že možno aj keď boli príbuzenské vzťahy v kmeni lovcov a zberačov dosť silné

na to, aby vytvorili selekčný tlak pomáhať aj nie priamo príbuzným jedincom, ak bolo nablízku *viacero* potenciálnych pomocníkov, mohli vzniknúť genetické preteky byť tým *posledným*, ktorý vykročí vpred. Každý otáľa, dúfajúc, že to urobí niekto iný. Na ľudstvo práve teraz čaká viacero hrozieb vyhynutia celého druhu, a musím vám povedať, že veľa ľudí nevykračuje z radu. Ak tento boj prehráme vďaka tomu, že sa prakticky nikto nedostavil na bojisko, potom – ako pravdepodobne mnoho druhov, ktoré dnes okolo seba nevidíme – sme sa vyvinuli k vyhynutiu.

Rakovinovým bunkám sa v tele darí pomerne dobre, vlastnia a hromadia stále viac prostriedkov, víťazia nad svojimi poslušnejšími kolegami. Na chvíľu.

Viacbunkové organizmy dokážu existovať iba vďaka tomu, že si vyvinuli silné vnútorné mechanizmy, ktoré *stavajú evolúciu mimo zákon*. Ak sa bunky začnú vyvíjať, rýchlo sa vyvinú k vyhynutiu: organizmus zomrie.

Nechváľte teda evolúciu za to, že sa stará o jednotlivca; takmer všetci vaši predkovia sú mŕtvi. Nechváľte evolúciu za to, že sa stará o druh; nikto ešte nenašiel zložitú adaptáciu, ktoré by sa dala vysvetliť iba ako ochraňujúca daný druh, a matematika naznačuje, že je to prakticky nemožné. Veru, je úplne možné, aby sa druh vyvinul k vyhynutiu. Ľudstvo možno práve teraz dokončuje tento proces. Nemôžete chváliť evolúciu dokonca ani za to, že sa stará o gény; bitka medzi dvoma alternatívnymi alelami o to isté miesto je hra o frekvenciu s nulovým súčtom.

Spôsobilosť nie je vždy váš kamarát.

*→

136. Tragédia skupinového selekcionizmu

Pred rokom 1966 nebolo nezvyčajné vidieť vážnych biológov obhajovať evolučné hypotézy, ktoré by sme dnes považovali za magické myslenie. Tieto zahmlené pojmy zohrali dôležitú historickú rolu vo vývoji neskoršej evolučnej teórie; boli to chyby, ktoré vyvolali opravy; tak ako pochabosť anglických kráľov spôsobila, že vznikla Magna Carta a konštitučná demokracia.

Ako príklad romantického myslenia, Vero Wynne-Edwards, Warder Allee, J. L. Brereton a ďalší verili, že dravce dobrovoľne obmedzia svoje rozmnožovanie, aby sa vyhli premnoženiu vo svojej lokalite a vyčerpaniu populácie koristi.

Lenže evolúcia neotvára stavidlá ľubovoľným účelom. Nemôžete vysvetliť štrkáčove štrkadlo povedaním, že existuje v prospech iných zvierat, ktoré by inak štrkáč uhryzol. Žiadna víla Evolúcia neurčuje zvonka, kedy by sa nejaký gén *mal* rozšíriť; účinok daného génu musí nejako *priamo spôsobiť*, že sa tento gén stane častejším v nasledujúcej generácii. Je jasné, prečo naše ľudské estetické cítenie pri pohľade na populačný pád líšok, ktoré zožrali všetky zajace, kričí: „Niečo by sa malo urobiť!“ Ale ako by komplex génov na *obmedzenie rozmnožovania* – zo všetkých vecí práve toto! - sám spôsobil, že bude častejší v nasledujúcej generácii?

Človek, ktorý by navrhoval nejakú malú hračkársku ekológiu – pre zábavu, ako keď sa modelujú železnice – by bol asi mrzutý, keby sa jeho starostlivo zostavené populácie líšok a zajacov sami zničili tým, že by líšky zožrali všetky zajace a potom by samotné skapali od hladu. Človek by sa teda babral s tou hračkárskou ekológiou – a človeku by ako prvé samozrejmé riešenie napadlo obmedziť rozmnožovanie líšok – dokiaľ by ekológia nevyzerala pekne a upravene. Príroda samozrejme nie je človek, ale *to* nás nezastaví – keď už vieme, čo *my* chceme z *estetického* hľadiska, musíme vymyslieť prijateľný argument, ktorý prírodu presvedčí, aby chcela *to isté* z *evolučných* dôvodov.

Samozrejme, výber na úrovni jednotlivca nespôsobí obmedzovanie sa jednotlivcov v rozmnožovaní. Jednotlivci, ktorí sa budú rozmnožovať bez obmedzenia, budú mať prirodzene viac potomkov než jednotlivci, ktorí sa obmedzovali.

(Výber na úrovni jednotlivca nevytvorí *jednotlivcove obetovanie príležitostí na rozmnožovanie*. Výber na úrovni jednotlivca môže však vytvoriť jednotlivcov, ktorí po získaní všetkých dostupných

→ http://lesswrong.com/lw/l5/evolving_to_extinction/

zdrojov použijú tieto zdroje na vytvorenie 4 veľkých vajec namiesto 8 malých vajec – *nie* kvôli šetreniu prostriedkov, ale pretože toto je spôsob, ako *jednotlivec* optimalizuje (počet vajec) × (pravdepodobnosť prežitia vajca). Toto však nerieši problém spoločných zdrojov.)

Predstavme si však, že sa populácia daného živočíšneho druhu rozdelí na subpopulácie, ktoré sú prevažne izolované a iba zriedkavo sa navzájom pária. Potom by iste subpopulácie, ktoré obmedzujú svoje rozmnožovanie, mali menšiu pravdepodobnosť vyhynúť, takže by vysielali viac poslov zakladajúcich nové kolónie na znovuosídlenie území vyhynutých populácií.

Problém tohto scenára nie je, že by bol matematicky *nemožný*. Problém je, že je *možný, ale veľmi ťažký*.

Zásadným problém je to, že výhody obmedzeného rozmnožovania sa neobmedzujú iba na tých, ktorí obmedzili svoje rozmnožovanie. Ak niektoré líšky obmedzia počet svojich mláďat žerúcich zajace, potom tieto nezožrané zajace nepripadnú *iba* tým mláďatám, ktoré nesú adaptáciu na obmedzenie rozmnožovania. Neobmedzujúce sa líšky a ich omnoho početnejšie mláďatá ochotne zožerú všetky zajace navyše. Jediný spôsob, ako môže obmedzujúci gén prežiť proti takémuto tlaku je, keď výhody obmedzovania sa pripadnú najmä tým, ktorí sa obmedzujú.

Konkrétne treba, aby $C/B < F_{ST}$, kde C sú náklady altruistu pre darcu, B je úžitok altruizmu pre príjemcu, a F_{ST} je priestorová štruktúra populácie: priemerná *príbuznosť* medzi náhodne vybraným organizmom a jeho náhodne vybraným susedom, kde „sused“ je každá ďalšia líška, ktorá bude mať úžitok z obmedzenia sa altruistickej líšky.[156]

Sú teda náklady obmedzeného rozmnožovania dostatočne malé a empirický zisk z menšieho hladu dostatočne veľký v porovnaní s empirickou priestorovou štruktúrou populácií líšok a populácií zajacov, aby tu mohol fungovať argument skupinového výberu?

Matematika naznačuje, že je to veľmi nepravdepodobné. Napríklad v tejto simulácii sú náklady altruistu 3 %, skupina čistých altruistov má spôsobilosť dvakrát väčšiu než skupina čistých sebcov, veľkosť subpopulácie je 25, a 20 % úmrtí sa nahrádza poslami z inej skupiny: výsledkom je rovnováha pre sebectvo a altruizmus. Ak veľkosť subpopulácie zdvojnásobíme na 50, zafixuje sa sebectvo; ak náklady na altruizmus zvýšime na 6 %, zafixuje sa sebectvo; ak výhody z altruizmu zmenšíme na polovicu, sebectvo sa zafixuje alebo získa veľkú väčšinu. Ak náklady na altruizmus prevýšia 10 %, musia byť susedské skupiny veľmi malé, iba okolo 5 členov, aby skupinová selekcia fungovala. Toto v prípade líšok obmedzujúcich svoje rozmnožovanie nevyzerá uveriteľne.

Teraz si už asi dokážete domyslieť, že zástancovia skupinového výberu nakoniec prehrali vedeckú debatu. Rozhodcom nebol matematický argument, ale empirické pozorovanie: líšky *neobmedzili* svoje rozmnožovanie (zabudol som, o ktoré živočíšne druhy sa jednalo; neboli to líšky a zajace) a ukázalo sa, že systémy dravcov a koriste sa rúcajú každú chvíľu. Neskôr skupinový výber opäť trochu ožil, v drasticky odlišnej podobe – matematicky povedané, susedská štruktúra *existuje*, z toho vyplýva nenulový tlak skupinového výberu, ktorý *nemusí* nutne dokázať prekonať protichodné tlaky individuálneho výberu, a ak toto nezahrniete do svojej matematickej rovnice, máte tam chybu, bodka. A vyvinuté mechanizmy donucovania (ktoré v pôvodnej teórii neboli) úplne menia hru. Prečo je teda táto dnes už historická vedecká debata vhodným materiálom pre *Overcoming Bias*?

Desaťročie po tejto kontroverzii mal jeden biológ fascinujúci nápad. Matematické podmienky aby skupinový výber prevážil nad individuálnym výberom boli príliš extrémne na to, aby sme ich našli v prírode. Prečo ich nevytvoriť umelo, v laboratóriu? Michael J. Wade urobil presne toto, opakovane vyberal populácie hmyzu na základe nízkeho počtu dospelých jednotlivcov v subpopulácii.[157] A čo bolo výsledkom? Začal hmyz obmedzovať svoje rozmnožovanie a žiť v tichom mieri s dostatkom jedla pre všetkých?

156 David Sloan Wilson, „A Theory of Group Selection,“ *Proceedings of the National Academy of Sciences of the United States of America* 72, no. 1 (1975): 143–146.

157 Michael J. Wade, „Group selections among laboratory populations of Tribolium,“ *Proceedings of the National Academy of Sciences of the United States of America* 73, no. 12 (1976): 4604–4607, doi:10.1073/pnas.73.12.4604.

Nie; dospelí sa prispôsobili na kanibalizovanie vajíčok a lariev, najmä samičích lariev.

Samozrejme, že výber početne malých subpopulácií nevyberal jednotlivcov, ktorí obmedzovali svoje *vlastné* rozmnožovanie; vyberal jednotlivcov, ktorí žrali *cudzie* deti. Najmä dievčatá.

Akonáhle máte v ruke tento pokusný výsledok – zo spätného pohľadu extrémne zrejmý – začne vám zrazu byť jasné, ako pôvodní zástancovia skupinového výberu dovoli romantizmu, ľudskému estetickému zmyslu, aby zahmlil ich predpovede o prírode.

Toto je učebnicový príklad nepovšimnutej tretej možnosti vychádzajúci z racionalizácie vopred určeného spodného riadku, ktorá vytvorila falošné zdôvodnenie a potom sa motivovane zastavila. Zástancovia skupinového výberu nezačali s jasnými prázdnymi hlavami, nenašli myšlienku skupinového výberu a nezačali ju *nestranne* rozvíjať smerom k pravdepodobnému výsledku. Začali s krásnou predstavou, ako populácie líšok dobrovoľne obmedzia svoje rozmnožovanie na mieru únosnú pre populáciu zajacov, príroda v dokonalej harmónii; potom hľadali dôvod, prečo by sa toto mohlo stať a prišli s nápadom skupinového výberu; potom, keďže už vedeli, ako *chcú*, aby dopadol výsledok skupinového výberu, nehľadali žiadne *menej* krásne a estetické adaptácie, ktoré by skupinový výber mohol priniesť s väčšou pravdepodobnosťou. Keby sa namiesto toho *naozaj* pokúsili pokojne a nestranne predpovedať výsledok vyberania početne malých subpopulácií odolných voči hladomoru, napadlo by im kanibalstvo detí druhých organizmov alebo nejaký podobne „hnusný“ výsledok – *dávno predtým*, než by si predstavili niečo také evolučne výstredné ako *dobrovoľné obmedzenie jednotlivca pri rozmnožovaní*!

Toto zároveň ilustruje pointu, ktorú som sa snažil urobiť v Einstenovej drzosti: Keď je priestor odpovedí veľký, takmer všetka skutočná práca sa týka podpory jednej možnej odpovede na úroveň, keď jej jednotlivo venujeme pozornosť. Ak hypotéze venujeme nezaslúženú pozornosť – ak váš zmysel pre estetiku navrhuje krásny spôsob, ako by príroda mala fungovať, ale napriek tomu prirodzený výber nezahŕňa vílu Evolúciu, ktorá by zdieľala vaše nadšenie – toto samotné môže spečatiť váš osud, pokiaľ si nedokážete dostatočne vyčistiť hlavu a začať odznovu.

Čisto teoreticky, aj najhlúpejší človek na svete môže povedať, že Slnko svieti, a preto ešte nenastane tma. Aj keby vám odpoveď navrhol šialenec pod vplyvom LSD, mali by ste dokázať nestranne spočítať indície za a proti, a ak treba, prestať tomu veriť.

Čisto prakticky boli zástancovia skupinového výberu odsúdení k skaze, pretože ich spodný riadok bol pôvodne navrhnutý ich zmyslom pre estetiku, zatiaľ čo spodný riadok prírody bol výsledkom prirodzeného výberu. Tieto dva procesy nemajú žiade principiálny dôvod, prečo by mala byť korelácia medzi ich výstupmi, a naozaj, nebola. Všetky dodatočné zúrivé argumenty to nezmenia.

Ak začnete od svojej vlastnej túžby, čo by príroda mala urobiť, zvážite pozorované dôvody, prečo príroda robí veci, a potom si racionalizujete extrémne presvedčivý argument prečo by príroda mala vytvoriť váš obľúbený výsledok zo svojich vlastných dôvodov, potom príroda, žiaľ, *stále* nebude počúvať. Vesmír nemá rozum a nie je ovplyvniteľný chytrým politickým presviedčaním. Môžete celý deň argumentovať, prečo by gravitácia mala radšej nechať tiecť vodu *hore* kopcom, ale voda bez ohľadu na to stále skončí na tom istom mieste. Je to ako keby vás vesmír jednoducho nepočúval. J. R. Molloy povedal: „Príroda je dokonalý fanatik, pretože je tvrdohlavo a netolerantne oddaná svojim predsudkom a absolútne odmieta ustúpiť aj tým najpresvedčivejším ľudským racionalizáciám.“

Často svojim priateľom odporúčam evolučnú biológiu, pretože táto moderná oblasť sa snaží cvičiť svojich študentov proti racionalizovaniu, chybe vyžadujúcej si dodatočnú opravu. Fyzici a elektroinžinieri sa nemusia strarostlivo cvičiť, aby sa vyhýbali antropomorfizácii elektrónov, pretože elektróny sa nesprávajú ako rozumné. Prirodzený výber vytvára účelnosť, ktorá je človeku cudzia, a študenti evolučnej teórie sú pred tým varovaní. Je to dobrý výcvik pre každého mysliaceho človeka, ale je *zvlášť* dôležitý, ak chcete jasne myslieť o iných cudzích mysliacich procesoch, ktoré nebudú fungovať rovnako ako vy.

*→

→ http://lesswrong.com/lw/kw/the_tragedy_of_group_selectionism/

137. Falošné optimalizačné kritériá

Už som predtým dosť dlho písal o rôznych formách racionalizácie kde naše názory vyzerajú byť zladené s indíciami omnoho silnejšie než v skutočnosti sú. Ale pritom to so zdôrazňovaním tohto bodu nepreháňam. Keby sme dokázali poraziť toto základné meta-skreslenie a vidieť, čo každá hypotéza *naozaj* predpovedá, dokázali by sme sa pozviechať z takmer každej inej faktickej chyby.

Teórii rozhodovania nastavíme zrkadlo tým, že sa pozrieme, ktorú možnosť *naozaj* podporuje výberové kritérium. Ak vaše hlásané morálne princípy žiadajú, aby ste poskytli laptopy pre všetkých, *naozaj* to podporuje nákup miliióndolárového drahokamami vykladaného laptopu pre seba, alebo vynaloženie rovnakého množstva peňazí na zaslanie 5000 OLPC?

Zdá sa, že máme vyvinutú zručnosť argumentovať, že prakticky hocijaký cieľ si vyžaduje prakticky hocijaký čin. Teoretický flogistonik vysvetľujúci, prečo magnézium pri horení získava hmotnosť, je nič oproti invizítorovi vysvetľujúcemu, prečo Božia nekonečná láska voči všetkým jeho deťom vyžaduje, aby niektoré z nich boli upálené na hranici.

V tomto nie je žiadne tajomstvo. Politika bola vlastnosťou pravekého prostredia. Sme potomkami tých, ktorí najpresvedčivejšie argumentovali, že dobro kmeňa vyžaduje, aby bol popravený ich nenávidený sok Uglak. (Určite nie sme potomkami Uglaka.)

A predsa... je možné *dokázať*, že keby Robertovi Mugabemu záležalo *iba* na tom, čo je dobré pre Zimbabwe, že by sa vzdal svojej funkcie prezidenta? Dokážete *argumentovať*, že by toto rozhodnutie vychádzalo z tohto cieľa, ale neukázali sme si práve, že ľudia dokážu spojiť hocijaký cieľ s hocijakým rozhodnutím? Odkiaľ viete, že vy máte pravdu a Mugabe sa mýli? (Na taký odhad by bolo pár dôvodov, ale počkajte chvíľku.)

Ľudské motívy sú mnohoraké a skryté, naše rozhodovanie procesy sú mimoriadne zložité, tak ako naše mozgy. A samotný svet je mimoriadne zložitý, pri každej voľbe pravidiel do skutočného sveta. Vieme vôbec *dokázať*, že ľudia racionalizujú – že systematicky skresľujeme spojenie medzi princípmi a rozhodnutiami – keď nám chýba jediné pevné miesto, na ktorom by sme mohli stáť? Keď neexistuje spôsob, ako *presne* zistiť, čo vyplýva hoci len z jediného optimalizačného kritéria? (V skutočnosti môžeme skrátka pozorovať, ako ľudia *nesúhlasia* s pracovnými problémami spôsobom, ktorý zvláštne koreluje s ich vlastnými záujmami, zatiaľ čo popierajú, že by takéto záujmy boli v hre. Ale opäť, chvíľku ešte počkajte.)

Kde je ten štandardizovaný, otvorený, všeobecne inteligentný, konsekvencialistický optimalizačný proces, do ktorého môžeme vložiť celú morálku vo formáte súboru XML, aby sme zistili, čo táto morálka *naozaj* odporúča, keď ju aplikujeme na náš svet? Existuje čo len jediný príklad zo skutočného života, kde vieme *presne*, čo dané kritérium výberu odporúča? Kde je ten *čistý* morálny mysliteľ – so známou funkciou úžitku, očistenou od všetkých ostatných bočných túžob, ktoré by mohli rozladiť jeho optimalizáciu – ktorého dôveryhodný výstup môžeme dať do kontrastu s ľudskými racionalizáciami tej istej funkcie úžitku?

No samozrejme náš starý známy, cudzí boh! Prirodzený výber je zaručene bez láskavosti, bez lásky, bez súcitu, bez estetických ohľadov, bez politického straníctva, bez ideologických spojencov, bez akademických ambícií, bez libertariánstva, bez socializmu, bez Modrých a bez Zelených. Prirodzený výber nevie *maximalizovať* podľa svojho kritéria celkovej genetickej spôsobilosti – až taký múdry nie je. Ale keď sa pozriete na výsledok prirodzeného výberu, zaručene pozeráte na výsledok, ktorý bol optimalizovaný *iba* na genetické prispôsobenie a nie na záujmy amerického poľnohospodárskeho rezortu.

V historických prípadoch evolučnej vedy – napríklad v Tragédii skupinového selekcionizmu – môžeme priamo porovnať ľudské racionalizácie s výsledkom *čistej* optimalizácie na známe kritérium. Čo si myslel Wynne-Edwards, že bude výsledkom skupinového výberu na malú veľkosť populácie? Dobrovoľné obmedzenie jednotlivcov pri rozmnožovaní a dosť jedla pre všetkých. Čo bol skutočný výsledok v laboratóriu? Kanibalizmus.

Teraz sa môžete opýtať: Sú tieto historické prípady evolučnej vedy naozaj relevantné pre ľudskú morálku, ktorej je úplne ukradnuté genetické prispôsobenie, pokiaľ sa postaví do cesty láske, súcitu,

estetike, zdraviu, slobode, spravodlivosti, a tak ďalej? Ľudská spoločnosť predsa do 20. storočia ani nepoznala pojem „celková genetická spôsobilosť“.

Ale ja sa na oplátku opýtam: Ak nedokážeme jasne vidieť výsledok jedného monotónneho optimalizačného kritéria – ak sa nedokážeme naučiť počuť jeden čistý tón – ako potom budeme počúvať celý orchester? Ako uvidíme, že: „Vždy buď sebec“ alebo: „Vždy poslúchaj vládu“ sú mizerné riadiace princípy pre ľudí – ak si myslíme, že ešte aj *optimalizácia génov na celkovú spôsobilosť* nám dá organizmy, ktoré obetujú svoje príležitosti na rozmnožovanie v mene uchovania spoločenských zdrojov?

Aby sme sa naučili vidieť jasne, potrebujeme jednoduché cvičné príklady.

*→

138. Vykonávatelia adaptácií, nie maximalizátori spôsobilosti

> Na jednotlivé organizmy je lepšie myslieť ako na vykonávateľov adaptácií než na maximalizátorov spôsobilosti.
>
> --John Tooby a Leda Cosmides, Psychologické základy kultúry[158]

Pred päťdesiatimi tisícmi rokov chuťové bunky *Homo sapiens* smerovali svojich nositeľov k najvzácnejším, najkritickejším potravinovým zdrojom – cukru a tuku. Iným slovom, kalórie. Dnes sa okolnosti zmenili, ale samotné chuťové bunky sa nezmenili. Kalórie zďaleka nie sú vzácne (vo vyspelých krajinách), ale aktívne škodia. Mikronutrienty, ktorých bolo spoľahlivý nadbytok v listoch a orechoch, chýbajú v chlebe, ale naše chuťové bunky sa nesťažujú. Kopček zmrzliny je superstimul, obsahuje viac cukru, tuku a soli než čokoľvek v pravekom prostredí.

Žiaden človek, ktorý by mal *vedomý* cieľ maximalizovať celkovú genetickú spôsobilosť svojich alel, by nezjedol koláč, dokiaľ by nehladoval. Lenže na jednotlivé organizmy je lepšie myslieť ako na vykonávateľov adaptácií než na maximalizátorov spôsobilosti.

Skrutkovač s hlavicou Phillips, hoci ho jeho dizajnér navrhol s cieľom otáčať skrutky, sa neprispôsobí sám od seba, aby zapadol do plochej skrutky. Vytvorili sme tieto nástroje, ale existujú nezávisle na nás, a pokračujú nezávisle na nás.

Atómy skrutkovača nemajú v sebe maličké značky XML, ktoré by opisovali ich „objektívny“ účel. Dizajnér tým niečo zamýšľal, to áno, ale to nie je to isté ako čo *sa stane* v skutočnom svete. Keby ste zabudli, že dizajnér je niečo iné ako nadizajnovaná vec, mohli by ste si myslieť: „*Účelom* skrutkovača je otáčať skrutky“ - akoby toto bolo vnútornou vlastnosťou samotného skutkovača a nie vlastnosťou dizajnérovho stavu mysle. Mohli by ste byť prekvapení, že sa skrutkovač sám neprestaval na skrutkovač s plochou hlavicou, keď predsa *účelom* skrutkovača je otáčať skrutky.

Príčinou existencie skrutkovača je dizajnérova myseľ, ktorá si predstavila imaginárnu skrutku a predstavila si otáčanie imaginárnou rúčkou. *Skutočné* fungovanie skrutkovača, jeho *skutočné* zapadnutie do hlavice skutočnej skrutky *nemôže* byť objektívnou príčinou existencie skrutkovača: Budúcnosť nemôže byť príčinou minulosti. Ale dizajnérov mozog, skutočná vec existujúca v minulosti, ten môže byť príčinou skrutkovača.

Dôsledky existencie skrutkovača nemusia zodpovedať imaginárnym dôsledkom v dizajnérovej mysli. Čepeľ skrutkovača sa môže šmyknúť a pozerať ruku používateľa.

A samotný *zmysel* skrutkovača – nuž to je niečo, čo existuje v hlave používateľa, nie v maličkých značkách na atómoch skrutkovača. Dizajnér mohol mať zámer, že bude otáčať skrutky. Vrah si ho mohol kúpiť, aby ho použil ako zbraň. A potom ho mohol náhodou stratiť, a zdvihlo ho dieťa, ktoré ho použilo ako dláto.

→ http://lesswrong.com/lw/kz/fake_optimization_criteria/

158 John Tooby and Leda Cosmides, „The Psychological Foundations of Culture,“ in *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*, ed. Jerome H. Barkow, Leda Cosmides, and John Tooby (New York: Oxford University Press, 1992), 19–136.

Takže *príčina* skrutkovača, jeho *tvar*, jeho *dôsledky*, a jeho rôzne *významy*, to sú všetko rôzne veci; a iba *jedna* z týchto vecí sa nachádza v skrutkovači samotnom.

Odkiaľ pochádzajú chuťové bunky? Nie od inteligentného dizajnéra, ktorý by si predstavoval ich dôsledky, ale zo zmrazenej histórie predkov: Adam mal rád sladké, zjedol jablko a rozmnožil sa; Barbara mala rada sladké, zjedla jablko a rozmnožila sa; Charlie mal rád sladké, zjedol jablko a rozmnožil sa; a o 2763 generácií neskôr sa alela zafixovala v populácii. Kvôli pohodlnejšiemu mysleniu občas túto veľkú históriu stlačíme a povieme: „Urobila to evolúcia.“ Ale to nie je rýchla lokálna udalosť, ako keď si človek dizajnér predstavuje skrutkovač. Toto je *objektívna príčina* chuťovej bunky.

Čo je *objektívny tvar* chuťovej bunky? Technicky povedané, je to molekulárny senzor pripojený na posilňujúci okruh. To pridáva ďalšiu nepriamu úroveň, pretože chuťová bunka priamo nezháňa potravu. Ovplyvňuje myseľ organizmu, spôsobuje, že organizmus chce jesť jedlá podobné tomu, ktoré práve zjedol.

Čo je *objektívny dôsledok* chuťovej bunky? U moderného človeka z vyspelej krajiny sa prejavuje vo viacerých reťaziach kauzality: od túžby jesť viac čokolády, po plán jesť viac čokolády, po pribranie, po menej randenia, po menej úspešné rozmnožovanie. Tento dôsledok je priamym *opakom* kľúčovej pravidelnosti v dlhej reťazi pravekých úspechov, ktoré spôsobili tvar chuťovej bunky. Lenže prejedanie sa stalo problémom iba pomerne nedávno, žiadna významná evolúcia (stlačená pravidelnosť predkov) už neovplyvnila tvar chuťovej bunky.

Aký je *zmysel* jedenia čokolády? To je medzi vami a vašou morálnou filozofiou. Ja osobne si myslím, že čokoláda chutí dobre, ale želám si, aby bola menej škodlivá; prijateľným riešením by mohla byť premena čokolády alebo premena mojej biochémie.

Pospájaním niekoľkým pojmov dokopy by sme mohli viacmenej povedať: „Moderní ľudia robia dnes to, čo by bolo šírilo ich gény v spoločnosti lovcov-zberačov, bez ohľadu na to, či to pomáha génom v modernej spoločnosti.“ Lenže ani toto nie je úplne správne, pretože my sa *v skutočnosti* nepýtame, ktoré správanie by maximalizovalo celkovú prispôsobenosť našich predkov. A mnohé naše dnešné činnnosti nemajú praveký ekvivalent. V spoločnosti lovcov-zberačov neexistovalo nič také ako čokoláda.

Preto je lepšie vidieť naše chuťové bunky ako *adaptáciu* prispôsobenú pravekým podmienkam, ktoré zahŕňali hladovanie a jablká a pečené zajace, ktorú moderní ľudia *vykonávajú* v novom kontexte, ktorý zahŕňa lacnú čokoládu a ustavičné bombardovanie reklamou.

Preto sa hovorí: Na jednotlivé organizmy je lepšie myslieť ako na vykonávateľov adaptácií než na maximalizátorov spôsobilosti.

*→

139. Evolučná psychológia

Povedať „rozmnožovací orgán“ je rovnako redundantné ako povedať „protokol TCP/IP“. *Všetky* orgány sú rozmnožovacie orgány. Odkiaľ si myslíte, že pochádzajú vtáčie krídla? Od víly Vtáčej Evolúcie, ktorá si myslí, že lietanie je pôvabné? Vtáčie krídla sú tu preto, lebo prispievali k rozmnožovaniu vtáčích predkov. Rovnako ako vtáčie srdce, pľúca, alebo pohlavné orgány. Nanajvýš môžeme považovať za vhodné rozlišovať medzi orgánmi, ktoré slúžia na rozmnožovanie *priamo* a *nepriamo*.

Toto pozorovanie platí aj pre mozog, najzložitejší systém orgánov známy v biológii. Niektoré orgány mozgu slúžia na rozmnožovanie priamo, napríklad žiadostivosť; iné slúžia rozmnožovaniu nepriamo, napríklad hnev.

Odkiaľ pochádza ľudský hnev? Od víly Ľudskej Evolúcie, ktorá si myslela, že hnev bude hodnotná vlastnosť? Nervové obvody hnevu sú rozmnožovacím orgánom rovnako isto ako vaša pečeň. *Homo sapiens* má hnev preto, lebo hnevajúci sa predkovia mali viac detí. *Neexistuje iný spôsob, ako sa tam mohol dostať.*/

→ http://lesswrong.com/lw/l0/adaptationexecuters_not_fitnessmaximizers/

Tento *historický* fakt o pôvode hnevu mätie priveľa ľudí. Hovoria: „Počkaj, ty hovoríš, že keď sa hnevám, podvedome sa snažím mať deti? To vôbec nie je to, na čo myslím, keď ma niekto udrie do nosa."

Nie. *Nie.* **Nie. NIE!**

Na jednotlivé organizmy je lepšie myslieť ako na vykonávateľov adaptácií než na maximalizátorov spôsobilosti. *Príčina* adaptácie, *tvar* adaptácie, a *dôsledky* adaptácie, sú všetko rôzne veci. Keby ste stavali hriankovač, nebudete očakávať, že zmení svoj tvar, keď sa doňho pokúsite napchať celý bochník; áno, vy ste chceli, aby robil hrianky, ale tento zámer je fakt o vás, nie fakt o hriankovači. Hriankovač nemá zmysel pre svoj vlastný účel.

Lenže hriankovač nie je objekt, ktorý by mal úmysly. Nie je to myseľ, takže nemáme sklon pripisovať mu ciele. Keď *my* vidíme hriankovač fungovať tak, ako má, nemyslíme si, že o tom ten hriankovač vie, pretože si nemyslíme, že si hriankovače niečo *vedia*.

Je to ako v tom starom teste, kde vás požiadajú, aby ste nahlas povedali farbu písmen slova „modrá" (napísaného červenou). Pokusným osobám trvá dlhšie menovať túto farbu, pretože musia oddeliť význam písmen od farby písmen. Nemali by ste rovnaký problém povedať farbu písmen slova „vietor" (napísaného zelenou).

Lenže ľudský mozog, okrem toho, že je to predmet historicky vytvorený evolúciou, je zároveň myseľ schopná mať svoje vlastné zámery, účely, túžby, ciele a plány. Aj včela aj človek sú dizajny, ale iba človek je dizajnér. Včela je ako „vietor", človek je ako „modrá".

Kognitívne príčiny sú *ontologicky odlišné* od evolučných príčin. Sú vyrobené z rôzneho materiálu. Kognitívne príčiny vznikajú z neurónov. Evolučné príčiny vznikajú z predkov.

Najzrejmejším príkladom kognitívnej príčiny je vedomé rozhodovanie, napríklad úmysel ísť do supermarketu alebo plán opiecť si hrianku. Ale aj emócia existuje fyzicky v mozgu, ako vlak nervových impulzov alebo oblak šíriacich sa hormónov. Podobne aj inštinkt alebo záblesk predstavivosti alebo letmo potlačená myšlienka; keby ste vedeli nasnímať mozog v troch rozmeroch a rozumeli by ste jeho kódu, dokázali by ste ich tam *vidieť*.

Aj podvedomé poznanie existuje fyzicky v mozgu. „Moc korumpuje," všimol si Lord Acton. Stalin možno o sebe veril, že je altruista a pracuje pre najväčšie dobro najväčšieho počtu ľudí; a možno nie. Ale zdá sa pravdepodobné, že niekde v Stalinovom mozgu existovali nervové obvody, ktoré príjemne posilňovali vykonávanie moci, a nervové obvody, ktoré zachytávali očakávania zvýšenia alebo zníženia moci. Keby v Stalinovom mozgu neexistovalo nič, čo by korelovalo s mocou – žiadne svetielko, ktoré by sa rozsvietilo pri politickom ovládaní a zhaslo by pri politickej slabosti – ako by potom Stalinov mozog mohol vedieť, že ho moc má skorumpovať?

Tlaky evolučného výberu sú *ontologicky odlišné* od biologických útvarov, ktoré vytvárajú. Evolučnou príčinou vtáčích krídel sú milióny vtáčích predkov, ktoré sa rozmnožovali častejšie než iní vtáčí predkovia, so štatistickou pravidelnosťou, za ktorú vďačili vlastníctvu postupne zlepšených krídel v porovnaní s ich konkurentmi. Tento gigantický historicko štatistický makrofakt stláčame do slov: „Urobila to evolúcia."

Prírodný výber je ontologicky odlišný od tvorov; evolúcia nie je malá chlpatá vec číhajúca v nepreskúmanom lese. Evolúcia je kauzálna, štatistická pravidelnosť v histórii rozmnožovania predkov.

A táto logika platí aj pre mozog. Evolúcia vytvorila krídla, ktoré mávajú, ale nerozumejú mávaniu. Vytvorila nohy, ktoré chodia, ale nerozumejú chodeniu. Evolúcia vytvorila kosti z vápnikových iónov, ale samotné kosti nemajú predstavu o sile, a už vôbec nie o celkovej genetickej spôsobilosti. A evolúcia vytvorila mozgy schopné dizajnovania; a predsa tieto mozgy nemali o nič väčšiu predstavu o evolúcii než má vták o aerodynamike. Až do 20. storočia žiaden ľudský mozog explicitne nereprezentoval zložitý abstraktný pojem *celkovej genetickej spôsobilosti*.

Keď nám povedia, že: „Evolučným účelom hnevu je zvýšiť celkovú genetickú spôsobilosť," máme sklon skĺznuť do: „*Účelom* hnevu je rozmnožovanie" a potom: „Kognitívnym účelom hnevu je rozmnožovanie." Nie! Štatistická pravidelnosť histórie predkov nie je *v mozgu*, ani len podvedome, rovnako ako dizajnérove úmysly dať si hrianku nie sú v hriankovači!

Myslieť si, že váš zabudovaný okruh pre hnev vyslovene stelesňuje túžbu rozmnožovať sa, je ako myslieť si, že vaša ruka stelesňuje myšlienkovú túžbu dvíhať veci.

Vaša ruka nie je celkom odrezaná od vašich myšlienkových túžob. V konkrétnom prípade môžete ovládať ohýbanie svojich prstov úkonom vôle. Ak sa zohnete a dvihnete mincu, tak toto môže predstavovať úkon vôle; ale nebol to tento úkon vôle, čo spôsobilo, že vaša ruka vôbec narástla.

Treba rozlišovať medzi jednorazovou udalosťou konkrétneho hnevu (hnev-1, hnev-2, hnev-3) a nervovými okruhmi, na ktorých je postavený hnev. Udalosť hnevu je kognitívna príčina a udalosť hnevu môže mať kognitívne príčiny, ale vaša vôľa nezabudovala do vášho mozgu obvody pre hnev.

Musíte teda rozlišovať medzi udalosťou hnevu, obvodmi pre hnev, komplexom génov, ktoré vybudovali tento nervový základ, a pravekým makrofaktom, ktorý vysvetľuje prítomnosť tohto komplexu génov.

Ak niekedy existoval odbor, ktorý naozaj *vyžadoval* extrémne rozoberanie detailov, je to evolučná psychológia.

Uvážte, moji čitatelia, tento úbohý a veselý príbeh: Muž a žena sa stretli v bare. Muža priťahovala jej hladká pleť a pevné prsia, ktoré by boli príznakmi plodnosti v pravekom prostredí, ale ktoré v tomto prípade spôsobuje make-up a podprsenka. To však toho muža netrápia; jemu sa skrátka páčilo, ako vyzerá. Jeho nervové obvody pre zisťovanie hladkej pleti nevedeli, že ich účelom je zisťovať plodnosť, rovnako ako atómy jeho ruky neobsahovali malé značky XML s nápisom „<účel>dvíhať veci</účel>“. Ženu priťahoval jeho sebavedomý úsmev a rozhodné správanie, príznaky vysokého postavenia, ktoré by v pravekom prostredí znamenali schopnosť zabezpečiť prostriedky deťom. Mala v úmysle použiť antikoncepciu, ale jej detektory sebavedomého úsmevu to nevedeli, rovnako ako hriankovač nevie, že jeho dizajnér si na ňom chcel urobiť hrianku. Ona sa filozoficky nezaujímala o zmysel tejto rebélie, pretože jej mozog bol kreacionista a dôrazne popieral existenciu evolúcie. On sa filozoficky nezaujímal o zmysel tejto rebélie, pretože si iba chcel zasúložiť. Išli do hotela a vyzliekli sa. On si nasadil kondóm, pretože nechcel mať deti, iba dopamínovo noradrenalínovú dávku zo sexu, ktorý spoľahlivo produkoval potomstvo pred 50 000 rokmi, keď bolo nemennou črtou pravekého prostredia, že kondómy neexistujú. Mali sex, osprchovali sa a rozišli sa. Hlavným objektívnym dôsledkom bolo, že sa bar a hotel a výrobca kondómov udržali v biznise; čo nebolo kognitívnym účelom v ich mysliach, a nemalo to prakticky nič spoločné s hlavnými štatistickými pravidelnosťami rozmnožovania pred 50 000 rokmi, ktoré vysvetľujú, ako získali gény, ktoré vybudovali ich mozgy, ktoré vykonali všetko toto správanie.

Aby ste správne uvažovali o evolučnej psychológii, musíte zároveň brať do úvahy mnoho zložitých abstraktných faktov, ktoré silno súvisia a pritom sú významne oddelené, a ani raz to nespojiť ani nepomiešať.

*→

140. Obzvlášť elegantný evolučne psychologický argument

> V kanadskej štúdii z roku 1989 žiadali dospelých, aby si predstavovali smrť detí rôzneho veku a odhadovali, ktorá smrť by vyvolala najväčší pocit straty u rodiča. Výsledky, nakreslené na grafe, ukazujú ako žiaľ rastie až po obdobie tesne pred dospievanie, a potom začne klesať. Keď túto krivku porovnali s krivkou ukazujúcou rozmnožovací potenciál počas životného cyklu (vypočítanou na základe kanadských demografických údajov), korelácia bola dosť silná. Ale omnoho silnejšia – v skutočnosti takmer dokonalá -- bola korelácia medzi krivkami žiaľu týchto moderných Kanaďanov a krivkou rozmnožovacieho potenciálu lovcov a zberačov, !Kungov v Afrike. Inými slovami, vzorec meniaceho sa žiaľu bol takmer presne taký, aký by predpovedal Darwinovec pri daných demografických skutočnostiach v pravekom prostredí.

→ http://lesswrong.com/lw/l1/evolutionary_psychology/

--Robert Wright, Morálne zviera, sumarizujúc Crawforda a kol.[159]

Prvá korelácia bola 0,64, druhá bola extrémne vysokých 0,92 (N = 221).

Najzrejmejšia *neelegancia* tejto štúdie, ako je opísaná, je že ju robili tak, že žiadali dospelých ľudí, aby si predstavovali rodičovský žiaľ, namiesto aby sa pýtali skutočných rodičov s deťmi daného veku. (Pravdepodobne by to stálo viac / mali by menej pokusných osôb.) Ale rozumiem tomu tak, že tieto výsledky dobre sedeli s údajmi z podrobnejších štúdií rodičovského žiaľu, ktoré hľadali iné korelácie (napríklad čistú koreláciu medzi rodičovským žiaľom a vekom dieťaťa).

Ale s horeuvedenou výhradou, všimnime si niektoré elegantné stránky tohto pokusu:

- Korelácia 0,92 (!). To môže znieť podozrivo vysoko -- dokáže evolúcia naozaj tak *presne naladiť*? -- kým si neuvedomíte, že tento selekčný tlak bol nielen dosť veľký na to, aby presne naladil rodičovský žiaľ, ale aby ho *v prvom rade vôbec od nuly vytvoril*.
- Ľudia, ktorí hovoria, že evolučná psychológia nerobí žiadne predpovede, sú (ironicky) iba obeťami syndrómu „nikto nevie, čo veda nevie“. Nebyť evolučnej psychológie, ani by vám *nenapadlo, že sa dá takýto pokus urobiť*.
- Tento pokus ukazuje tak krásne a tak jasne, ako som len kedy videl, rozdiel medzi *vedomým alebo nevedomým motívom* a *vykonávaním adaptácie, ktorá nemá žiadnu citlivosť v reálnom čase na pôvodný selekčný tlak, ktorý ju vytvoril*.

Rodičovský žiaľ sa *ani len podvedome* netýka rozmnožovacej hodnoty -- inak by sa prispôsobil rozmnožovacej hodnote Kanaďanov namiesto rozmnožovacej hodnoty !Kungov. Žiaľ je adaptácia, ktorá dnes jednoducho existuje, je skutočná v mysli, a pokračuje svojou vlastnou zotrvačnosťou.

Rodičia sa starajú o svoje deti *nie* kvôli ich rozmnožovaciemu príspevku. Rodičia sa starajú o svoje deti kvôli deťom samotným; a *nekognitívny, evolučne historický* dôvod, prečo takéto mysle vôbec vo vesmíre existujú, je že deti nesú gény svojich rodičov.

Veru, evolúcia je ten dôvod, prečo vo vesmíre vôbec nejaké mysle existujú. Môžete teda vidieť, prečo chcem nakresliť ostrú čiaru okolo môjho cynizmu o skrytých motívoch na evolučne kognitívnej hranici; inak by som sa rovnako dobre mohlo postaviť do radu pri pokladni v supermarkete a povedať: „Hej! Vy počas blokovania mojich potravín správne spracovávate zrakové informácie iba preto, aby ste maximalizovali svoju celkovú genetickú spôsobilosť!“

1. Myslím si, že 0,92 je najvyššia korelácia, akú som kedy videl v evolučne psychologickom pokuse, a vôbec jedna z najvyšších korelácií, aké som kedy videl v psychologickom pokuse. (Aj keď som už videl napríklad udávať koreláciu 0,98 pri pýtaní sa jednej skupiny pokusných osôb: „Ako veľmi sa A podobá na B?“ a druhej skupiny: „Aká je pravdepodobnosť A za predpokladu B?“ pri otázkach typu: „Aká je pravdepodobnosť, že vyberiete 60 červených loptičiek a 40 bielych loptičiek z tohto suda, kde je 800 červených loptičiek a 200 bielych loptičiek?“ -- inými slovami to jednoducho spracovali akoby to bola tá istá otázka.)

Keďže tu sme všetci Bayesovci, môžeme zohľadniť svoje pôvodné predpoklady a opýtať sa, či aspoň *niečo* z tejto nečakane vysokej korelácie nie je otázkou šťastia. Evolučné presné ladenie môžeme asi brať ako dané; toto je veľký selekčný tlak, o čom tu hovoríme. Zvyšné zdroje podozrivo nízkej variancie sú (a) či si veľká skupina dospelých ľudí v priemere dokáže správne predstaviť stupne rodičovského žiaľu (zdá sa, že áno). A (b), či sú dnešní !Kungovia v tomto ohľade *typickými* pravekými lovcami a zberačmi, alebo či sú variácie *medzi* rôznymi typmi kneňov lovcov a zberačov príliš vysoké na to, aby dovolili koreláciu 0,92.

Ale aj keď zohľadníme takéto skeptické pôvodné predpoklady, korelácie 0,92 a N = 221 je dosť silná indícia, a naše výsledné postoje by mali byť vo všetkých týchto veciach menej skeptické.

2. Môžete považovať za neelegantnosť pokusu, že sa robil na základe *predstavy* fiktívneho žiaľu, namiesto spomienky na skutočný žiaľ. Ale je to práve táto *predstava* fiktívneho žiaľu, ktorá usmerňuje rodičovské správanie, aby *neprišli* o svoje dieťa! Z pohľadu evolúcie, dieťa, ktoré už naozaj zomrelo, je

159 Robert Wright, *The Moral Animal: Why We Are the Way We Are: The New Science of Evolutionary Psychology* (Pantheon Books, 1994); Charles B. Crawford, Brenda E. Salter, and Kerry L. Jang, „Human Grief: Is Its Intensity Related to the Reproductive Value of the Deceased?,“ Ethology and Sociobiology 10, no. 4 (1989): 297–307.

utopený náklad; evolúcia „chce“, aby sa rodič z tejto bolesti poučil, aby to druhýkrát neurobil, aby sa vrátil do svojho hedonického pevného bodu, a aby vychovával ďalšie deti.

3. Podobne, graf, ktorý koreluje s rodičovským žiaľom, sa týka *budúceho rozmnožovacieho potenciálu dieťaťa, ktoré sa dožilo daného veku*, a nie *súčtu utopených nákladov vychovávania dieťaťa, ktoré sa dožilo daného veku*. (Možno by sme dostali ešte *vyššiu* koreláciu, keby sme sa pokúsili zohľadniť alternatívne náklady na vychovanie dieťa od veku X do nezávislej dospelosti, a ignorovali všetky utopené náklady na výchovu dieťaťa do veku X?)

Ľudia si zvyčajne všímajú utopené náklady -- dá sa predpokladať, že je to buď adaptácia, ktorá nám má brániť príliš rýchlo striedať stratégie (kompenzuje za príliš horlivé všímanie si príležitostí?) alebo nešťastný pozostatok bolesti, ktorú cítime pri strate zdrojov.

Na druhej strane, evolúcia -- ani nejde o to, že by evolúciu „nezaujímali utopené náklady“, ale že evolúcia ani približne „nerozmýšľa“ týmto spôsobom; „evolúcia“ je jednoducho makrofakt o skutočných historických dôsledkoch na rozmnožovanie.

Takže -- samozrejme -- je adaptácia rodičovského žiaľu presne naladená spôsobom, ktorý nemá nič spoločné s minulými investíciami do dieťaťa, a všetko s budúcimi rozmnožovacími dôsledkami straty tohto dieťaťa. Prirodzený výber nie je posadnutý utopenými nákladmi tak ako vy.

Ale -- samozrejme -- adaptácia rodičovského žiaľu naďalej funguje tak, ako keby daný rodič žil v kmenu !Kungov a nie v Kanade. Väčšina ľudí by si tento rozdiel všimla.

Ľuďom a prirodzenému výberu šibe *rôznymi stabilnými komplikovanými spôsobmi*.

*→

141. Superstimuly a kolaps západnej civilizácie

Prinajmenšom traja ľudia zomreli následkom hrania online hier celé dni bez prestávky. Ľudia stratili svojich partnerov, zamestnania, deti kvôli hre World of Warcraft. Ak majú ľudia právo hrať počítačové hry – a je ťažké predstaviť si nejaké základnejšie právo – trh bude reagovať tým, že im ponúkne tie *najpútavejšie* počítačové hry, aké sa len dajú predať, až do bodu, kde sa mimoriadne pripútaní zákazníci odstránia z genofondu.

Ako sa tovar stane takým *neodolateľným*, že po 57 hodinách používania daného tovaru by zákazník radšej používal tento tovar ešte jednu hodinu, než by sa najedol alebo vyspal? (Niekto by azda mohol argumentovať, že zákazník urobil racionálne rozhodnutie, že bude radšej hrať Starcraft nasledujúcu hodinu, než žiť zvyšok svojho života, ale radšej tam nezachádzajme. Prosím.)

Sladká tyčinka je *superstimul*: obsahuje viac koncentrovaného cukru, soli a tuku než čokoľvek, čo existovalo v pravekom prostredí. Sladká tyčinka pôsobí na chuťové poháriky, ktoré sa vyvinuli v prostredí lovcov a zberačov, ale pôsobí na tieto chuťové poháriky omnoho silnejšie než čokoľvek, čo *v prostredí* lovcov a zberačov naozaj existovalo. Signál, ktorý kedysi spoľahlivo koreloval so zdravým jedlom, bol unesený, premožený bodom v priestore chutí, ktorý sa v pôvodnej tréningovej sade nenachádzal – neskutočne vzdialenou výnimkou z pravekých grafov. Chutnosť, ktorá kedysi predstavovala evolučne identifikovaný korelát zdravia, bola hacknutá dokonale prispôsobenou umelou látkou. Nanešťastie neexistuje žiadna rovnako silná trhová motivácia urobiť výsledné jedlo také zdravé ako je chutné. Predsa len, zdravosť ochutnať nevieme.

Známe video Dove Evolution ukazuje dôkladnú konštrukciu ďalšieho superstimulu: bežná žena premenená pomocou make-upu, starostlivého fotografovania, a nakoniec rozsiahlej úpravy vo Photoshope na billboardovú modelku – krásu nedosiahnuteľnú ľudskými ženami v neretušovanom skutočnom svete. Skutočné ženy sa zabíjajú (napr. supermodelky užívajú kokaín, aby si udržali nízku hmotnosť), aby udržali krok s konkurentkami, ktoré doslova neexistujú.

A rovnako môže byť počítačová hra omnoho *pútavejšia* než púha skutočnosť, dokonca aj cez jednoduchú počítačovú obrazovku, že niekto ju bude hrať bez jedla a bez spánku, kým doslova

→ http://lesswrong.com/lw/yj/an_especially_elegant_evpsych_experiment/

nezomrie. Nepoznám všetky triky používané v počítačových hrách, ale pár z nich viem uhádnuť – náročnosť umiestnená v kritickom bode medzi ľahkou a nemožnou, okamžité odmeňovanie, spätná väzba zobrazujúca stále rastúce skóre, spoločenské zapojenie v masívnych multiplayeroch.

Je nejaká hranica trhovej motivácie robiť počítačové hry ešte pútavejšími? Mohli by ste dúfať, že motivácia nebude existovať aspoň za tým bodom, kde hráči stratia svoje zamestnanie; predsa len, musia nejako zaplatiť svoje predplatné. To by naznačovalo, že existuje „ideálny stupeň" návykovosti hry, kde sa stred Gaussovej krivky zabáva, a iba pár nešťastníkov na jej chvoste sa stane natoľko závislými, že prídu o prácu. V roku 2007 nepretržité hranie World of Warcraft 58 hodín až do doslovnej smrti *je* stále ešte výnimkou, nie pravidlom. Lenže výrobcovia počítačových hier súťažia voči sebe navzájom, a ak viete svoju hru urobiť o 5% návykovejšou, môžete vďaka tomu prebrať 50% zákazníkov konkurencie. Viete si predstaviť, ako sa tento problém môže ešte *výrazne* zhoršiť.

Ak majú ľudia právo byť pokúšaní – a o tomto je celá slobodná vôľa – trh bude reagovať tým, že im ponúkne toľko pokušenia, koľko sa len dá predať. Je tu motivácia urobiť svoje podnety o 5% silnejším pokušením než sú podnety najvýznamnejších konkurentov. To pokračuje ďaleko za bod, v ktorom sa podnety z pohľadu praveku stali nenormálnymi superstimulmi. Zvážte, ako sa naše štandardy ženskej krásy potrebnej na predaj tovaru zvýšili v porovnaní s reklamami z 50-tych rokov. A ako ukazuje sladká tyčinka, trhová motivácia tiež pokračuje ďaleko za bod, kde superstimul začne zákazníkovi spôsobovať škody.

Prečo teda jednoducho nepovieme nie? Základným predpokladom trhovej ekonomiky je predsa, že v neprítomnosti násilia a podvodu sa ľudia vždy môžu odmietnuť zúčastniť škodlivej transakcie. (Do tej miery, do akej toto platí, je trhový mechanizmus nielen celkovo *najlepším* riešením, ale riešením s málo nevýhodami, ak vôbec nejakými.)

Organizmus, ktorý sa dlhodobo vzdáva potravy, zomrie, ako niektorí hráči počítačových hier zistili na vlastnej koži. V niektorých prípadoch v pravekom prostredí však zvyčajne užitočné (a preto lákavé) konanie môže byť v skutočnosti škodlivým. Ľudia majú v porovnaní s inými organizmami nezvyčajne silnú schopnosť vnímať tieto špeciálne prípady pomocou abstraktného myslenia. Na druhej strane máme aj sklon predstavovať si veľa následkov špeciálnych prípadov tam, kde žiadne nie sú, napríklad keď nám duchovia predkov zakazujú jesť úplne zdravé zajace.

Zdá sa, že evolúcia dosiahla kompromis, alebo možno len prihodila nové systémy navrch tých starých. *Homo sapiens* je stále pokúšaný jedlom, ale naša prerastená prefrontálna kôra nám dáva *obmedzenú* schopnosť odolávať pokušeniu. Nie neobmedzenú schopnosť – naši predkovia s priveľkou silou vôle sa možno sami vyhladovali na smrť ako obeť bohom, alebo sa príliš často vyvarovali cudzoložstva. Hráči počítačových hier, ktorí zomreli, museli (v istom zmysle) vynakladať silu vôle, aby vydržali hrať tak dlho bez jedla a spánku; evolučné riziko sebaovládania.

Odolať nejakému pokušeniu si vyžaduje vedomé vynaloženie vyčerpateľného zdroja duševnej energie. V skutočnosti nie je *pravda*, že sa dá „jednoducho povedať nie" – nie je *jednoduché* povedať nie, lebo nás to niečo stojí. Aj ľudia, ktorí v lotérii narodenia vyhrali silu vôle alebo prezieravosť, stále platia cenu za odolanie pokušeniu. Akurát sa im tá cena platí ľahšie.

Naša obmedzená sila vôle sa vyvinula, aby si poradila s pravekými pokušeniami; nemusí dobre fungovať proti dnešným lákadlám presahujúcim všetko, čo lovci a zberači poznali. Ešte aj tam, kde úspešne odolávame superstimulu, zdá sa možné, že toto vynaložené úsilie vyčerpáva silu vôle omnoho rýchlejšie než odolávanie pravekým pokušeniam.

Je verejné zobrazovanie superstimulov negatívnou externalitou ešte aj pre ľudí, ktorí hovoria nie? Mali by sme zakázať reklamy na čokoládové tyčinky, alebo obchody, ktoré otvorene hovoria „zmrzlina"?

Samotný fakt, že problém existuje, ešte neznamená (bez ďalšieho zdôvodnenia a značného bremena dôkazu), že ho vláda dokáže napraviť. Regulátorova kariérna motivácia sa nesústreďuje na výrobky, ktoré spájajú nízku škodu zákazníkovi s návykovými superstimulmi; sústreďuje sa na výrobky, ktoré majú šancu zlyhať spôsobom, ktorý zaujme pozornosť novín. A naopak, samotný fakt, že vláda niečo *nemusí* vedieť napraviť, ešte neznamená, že to *nedopadne* zle.

Na záver vám ponúknem argument fiktívnou indíciou: Internetový román Simona Funka After Life zobrazuje (okrem iných zápletiek) plánované vyhubenie biologických *Homo sapiens* – nie armádami pochodujúcich robotov, ale umelými deťmi, ktoré sú omnoho krajšie a milšie a zábavnejšie vychovávať než skutočné deti. Možno vyspelé spoločnosti demograficky kolabujú preto, lebo trh ponúka stále lákavejšie alternatívy oproti deťom, zatiaľ čo atraktivita vymieňania plienok zostáva konštantná. Kde sú reklamné billboardy s nápisom: „ROZMNOŽUJTE SA“? Kto zaplatí profesionálnym konzultantom imidžu, aby hádanie sa s mrzutými puberťákmi zobrazili lákavejšie než dovolenka na Tahiti?

„Nakoniec,“ píše Simon Funk, „bola ľudská rasa jednoducho marketingom vytlačená z existencie.“

*→

142. Ty si úlomok boha

Pred 20. storočím nemal žiaden človek explicitnú predstavu o „celkovej genetickej spôsobilosti“, jedinej a absolútnej posadnutosti slepého hlúpeho boha. Nemáme žiaden inštinktívny odpor voči kondómom alebo orálnemu sexu. Naše mozgy, tieto vrcholné rozmnožovacie orgány, nekontrolujú reprodukčnú efektívnosť predtým ako nám dajú sexuálne potešenie.

Prečo nie? Prečo *nie sme* vedome posadnutí celkovou genetickou spôsobilosťou? Prečo víla Ľudská Evolúcia vytvorila mozgy, ktoré dokážu vymyslieť kondóm? „Bolo by to také *ľahké*,“ pomyslí si človek, ktorý dokáže nadizajnovať nové zložité systémy za popoludnie.

Víla Evolúcia, ako všetci vieme, je posadnutá celkovou genetickou spôsobilosťou. Keď sa rozhoduje, ktoré gény povýši na všeobecné, neberie pritom do úvahy *nič* okrem počtu kópií, ktoré daný gén vytvorí. (Aké zvláštne!)

Ale ak je tvorca inteligencie takto posadnutý, prečo nevytvoril inteligentných činiteľov – nemôžeme ich nazvať ľuďmi – ktorí by sa takisto starali iba o celkovú genetickú spôsobilosť? Takí činitelia by mali sex iba ako prostriedok na rozmnožovanie a neobťažovali by sa sexom, ktorí by zahŕňal antikoncepciu. Jedli by z jasne zdôvodneného presvedčenia, že jedlo je potrebné na rozmnožovanie, nie preto, že by im chutilo, takže by nejedli cukríky, keby sa stali škodlivými pre prežitie alebo rozmnožovanie. Ženy po menopauze by sa starali o vnúčatá, dokiaľ by neochoreli natoľko, že by sa stali celkovou stratou zdrojov, a vtedy by spáchali samovraždu.

Toto vyzerá ako celkom jasné zlepšenie dizajnu – z pohľadu víly Evolúcie.

Teraz je jasné, že je ťažké vytvoriť dostatočne mocného konsekvencialistu. Prirodzený výber svojím spôsobom uvažuje konsekvenciálne, ale iba v tom, že závisí od *skutočných dôsledkov*. Ľudskí evoluční teoretici musia robiť pompézne abstraktné úvahy, aby si *predstavili* spojenie medzi adaptáciami a úspechom v rozmnožovaní.

Ale ľudské mozgy si jasne *dokážu* predstaviť tieto spojenia pomocou bielkoviny. Keď teda víla Evolúcia vyrobila ľudí, prečo sa unúvala dať im *nejakú* motiváciu okrem celkovej genetickej spôsobilosti?

Je to menej ako dvesto rokov odkedy mozog z bielkoviny prvýkrát pochopil pojem prirodzeného výberu. Moderný pojem „celkovej genetickej spôsobilosti“ je ešte jemnejší, vysoko abstraktný pojem. Na čom záleží, nie je počet spoločných génov. Šimpanzy s nami majú spoločných 95% génov. Na čom záleží, je spoločný genetický *rozptyl v rámci* reprodukujúcej sa populácie – vaša sestra je vám napoly príbuzná, pretože každú variáciu vo vašom genóme, v rámci ľudského druhu, má na 50 % aj vaša sestra.

Iba v poslednom storočí – dalo by sa tvrdiť, že iba za posledných päťdesiat rokov – začali evoluční biológovia naozaj rozumieť celému rozsahu príčin reprodukčného úspechu, veci ako vzájomný altruizmus a drahá signalizácia. Bez všetkých týchto veľmi detailných poznatkov by inteligentný činiteľ odhodlaný „maximalizovať celkovú spôsobilosť“ padol na hubu.

Prečo teda tieto vedomosti nenaprogramovať priamo do mozgov z bielkoviny? Prečo sme nedostali pojem „celkovej genetickej spôsobilosti“ priamo *v programe* spolu s knižnicou podrobných stratégií?

→ http://www.lesswrong.com/lw/h3/superstimuli_and_the_collapse_of_western/

Potom by ste mohli prideliť všetky tie posilňovače. Organizmus by sa narodil s vedomosťou, že tučné jedlá s vysokou pravdepodobnosťou povedú k spôsobilosti. Keby sa organizmus neskôr naučil, že toto už neplatí, prestal by jesť tučné jedlá. Mohli by ste refaktorovať celý systém. A potom by nevymyslel kondómy ani koláče.

Toto vyzerá, že by to v princípe celkom išlo. Občas natrafím na ľudí, ktorí celkom nechápu konsekvencializmus a povedia: „Ale keby organizmus nemal samostatnú túžbu jesť, vyhladoval by sa, a potom by sa nereprodukoval.“ Dokiaľ organizmus *o samotnom tomto fakte* vie, a má funkciu úžitku, ktorá si cení reprodukciu, automaticky bude jesť. V skutočnosti *presne* takéto konsekvencialistické uvažovanie použil samotný prirodzený výber, aby postavil automatických jedákov.

A čo zvedavosť? Nebol by konsekvencialista zvedavý iba vtedy, keby videl nejaký konkrétny dôvod byť zvedavý? A nespôsobilo by to, že by minul mnoho dôležitých vedomostí, ktoré prišli bez konkrétneho dôvodu na preskúmanie? Opäť, konsekvencialista by veci skúmal, keby si uvedomoval samotný tento fakt. Ak si vezmete pud zvedavosti u ľudí – ktorý nie je nerozlišujúci, ale reaguje na konkrétne vlastnosti problémov – táto zložitá adaptácia je čisto výsledkom konsekvencialistického uvažovania DNA, *nevyslovená* reprezentácia poznania: Predkovia, ktorí sa venovali tomuto druhu vypytovania, zanechali viac potomkov.

Takže v princípe je čistý reproduktívny konsekvencialista možný. V princípe možno všetku históriu predkov *nevyslovene* reprezentovanú kognitívnymi adaptáciami premeniť na *výslovne* reprezentovanú vedomosť, ktorou sa bude riadiť čistý konsekvencialista.

Ale slepý hlúpy boh nie je taký múdry. Evolúcia nie je ako človek programátor, ktorý môže zároveň refaktorovať celú architektúru kódu. Evolúcia nie je ako človek programátor, ktorý sú dokáže sadnúť a písať príkazy rýchlosťou šesťdesiat slov za minútu.

Milióny rokov pred konsekvencializmom hominidov existovalo učenie posilňovaním. Signály odmeny boli udalosti, ktoré spoľahlivo korelovali s rozmnožovaním. Nemohli ste žiadať mozog nehominida, aby predvídal, že dieťa, ktoré bude teraz jesť tučnejšie jedlo, prežije zimu. Preto DNA postavila bielkovinový mozog, ktorý vytvára signál odmeny pri jedení tučného jedla. Potom už je na danom organizme, aby sa naučil, ktorá korisť je najchutnejšia.

DNA zostavuje bielkovinové mozgy so signálmi odmeny, ktoré z *dlhodobého* hľadiska korelujú so spôsobilosťou rozmnožovať sa, ale z *krátkodobého* hľadiska korelujú so správaním organizmu. Nemusíte dôjsť na to, že jesť sladké ovocie na jeseň povedie k tráveniu kalórií, ktoré možno uložiť vo forme tuku, čo vám pomôže prežiť zimu, takže sa na jar budete môcť páriť a produkovať potomstvo. Jablko jednoducho chutí dobre, a váš mozog má za úlohu iba vymyslieť, ako dostať zo stromu čo najviac jabĺk.

A tak si organizmy vyvinú odmeny za jedenie, za budovanie hniezd, za odstrašenie súperov, za pomáhanie súrodencom, za objavovanie dôležitých právd, za vytváranie silných spojenectiev, za presvedčivé argumentovanie, a samozrejme za sex...

Keď sa objavili mozgy homonidov schopné uvažovať konsekvencialisticky vo viacerých oblastiach, uvažovali konsekvencialisticky o tom, ako získať *existujúce* posilnenia. Bol to pomerne jednoduchý trik, omnoho jednoduchší než postaviť celého „maximalizátora celkovej spôsobilosti“ od začiatku. Bielkovinové mozgy plánovali, ako získať kalórie a sex, bez nejakej výslovnej myšlienkovej predstavy „celkovej spôsobilosti“.

Človek inžinier by povedal: „Fíha, práve som vymyslel konsekvencialistu! Teraz môžem vziať všetky svoje doterajšie ťažko získané vedomosti o tom, ktoré správanie zlepšuje spôsobilosť, a zadať ich výslovne! Môžem premeniť celé toto zložité zariadenie na učenie pomocou posilňovania na jednoduché oznámenie vedomosti, že ‚tučné jedlá a sex zvyčajne zlepšujú tvoju celkovú spôsobilosť‘. Konsekvencialistické uvažovanie sa už automaticky postará o zvyšok. Plus, nebude jasne zlyhávať tým, že vymyslí kondómy!“

No ale človek inžinier by takisto ani nenavrhol sietnicu naopak.

Slepý hlúpy boh nie je nejaký jednotný cieľ, ale mnohonásobne rozdrobená pozornosť. Líšky sa vyvíjajú, aby chytali zajace, zajace sa vyvíjajú, aby unikli líškam; evolúcií je toľko, koľko je druhov.

Ale *vnútri* každého z týchto druhov je slepý hlúpy boh posadnutý iba celkovou genetickou spôsobilosťou. Necení si žiadnu inú vlastnosť, dokonca ani prežitie, pokiaľ nezvyšuje reprodukčnú spôsobilosť. Nemá zmysel, aby mal organizmus oceľovú kožu, ak má v dôsledku toho o 1 % nižšiu reprodukčnú kapacitu.

Napriek tomu, keď slepý hlúpy boh vytvoril bielkovinové počítače, jeho monomaniakálne zameranie na celkovú genetickú spôsobilosť sa verne neprenieslo. Jeho optimalizačné kritériá sa úspešne nekvajnovali. My, dielo evolúcie, sme evolúcii rovnako cudzí ako je náš Stvoriteľ cudzí nám. Jedna čistá funkcia úžitku sa rozdrobila na tisíce zlomkov túžby.

Prečo? V prvom rade preto, lebo evolúcia je absolútne hlúpa. Ale tiež preto, lebo *prvé* bielkovinové počítače neboli zďaleka také *všeobecné* ako slepý hlúpy boh, a dokázali spracovávať iba krátkodobé túžby.

V závere našej analýzy, pýtať sa, prečo evolúcia nevytvorila ľudí, aby maximalizovali celkovú genetickú spôsobilosť, je ako pýtať sa, prečo evolúcia nedala ľuďom ribozóm a nepovedala im, aby si navrhli svoju vlastnú biochémiu. Pretože evolúcia nedokáže refaktorovať kód tak rýchlo, to je dôvod. Ale možno keby prirodzený výber pokračoval ešte miliardu rokov, stalo *by* sa presne toto, keby bola inteligencia natoľko hlúpa, že by dovolila hlúpemu bohu pokračovať v jeho vláde.

Trieska v Božom oku od Nivena a Pournella vykresľuje inteligentný živočíšny druh, ktorý zostal biologickým príliš dlho, pomaly bol naozaj zotročený evolúciou, postupne sa premenil na skutočných maximalizátorov spôsobilosti posadnutých rozmnožovaním sa rýchlejšie než druhí. Ale našťastie sa toto nestalo. Nie na Zemi. Aspoň zatiaľ nie.

Ľudia teda majú radi chuť cukru a tuku, máme radi svojich synov a dcéry. Hľadáme sociálne postavenie a sex. Spievame, tancujeme, hráme sa. Učíme sa, pretože máme radi učenie.

Tisíc jemných chutí zodpovedajúcich pravekým posilňovačom, ktoré kedysi korelovali s reprodukčnou spôsobilosťou – dnes vyhľadávaných bez ohľadu na to, či pomáhajú rozmnožovaniu. Sex s antikoncepciou, čokoláda, hudba dávno mŕtveho Bacha na CD.

A keď sa nakoniec dozvieme o evolúcii, pomyslíme si: „Byť celý deň posadnutý celkovou genetickou spôsobilosťou? Čo je na *tomto* zábavné?“

Jediný monomaniakálny cieľ slepého hlúpeho boha sa rozdrobil na tisíce zlomkov túžby. A to je v poriadku, myslím si, hoci to hovorím ako človek. Lebo inak, čo by sme robili s budúcnosťou? Čo by sme robili s milardou galaxií na nočnej oblohe? Zaplnili ich maximálne efektívnymi replikátormi? Mali by byť naši potomkovia vedome posadnutí maximalizovaním svojej celkovej genetickej spôsobilosti, vnímať všetko ostatné iba ako prostriedok na tento účel?

Byť tisícom zlomkov túžby nie je vždy zábava, ale prinajmenšom to nie je *nuda*. Niekde po ceste sme si vyvinuli chute na novinky, zložitosť, eleganciu a výzvu – chute, ktoré posudzujú monomaniakálne zameranie slepého hlúpeho boha a hodnotia ho ako esteticky neuspokojivé.

A áno, aj samotné tieto chute sme dostali zo zlomkov slepého hlúpeho boha. No a?

* →

—

→ http://lesswrong.com/lw/l3/thou_art_godshatter/

M: Krehké ciele

143. Viera v inteligenciu

Neviem, ako by Garry Kasparov ťahal v šachovej hre. Čo je potom empirickým obsahom môjho názoru, že „Kasparov je vysoko inteligentný šachista“? Aké vnemy v reálnom svete mi môj názor prikazuje očakávať? Je to iba chytro maskovaná forma úplnej nevedomosti?

Aby som dilemu vyostril, predstavme si, že Kasparov hrá proti nejakému iba šachovému veľmajstrovi, pánovi G., ktorý sa nedostal do svetového šampionátu. Moja vlastná schopnosť je príliš slabá na to, aby som rozoznal tieto úrovne šachovej zručnosti. Keď sa pokúšam uhádnuť ďalší Kasparovov ťah, alebo ďalší ťah pána G., dokážem sa akurát pokúšať uhádnuť „najlepší šachový ťah“ pomocou mojich skromných znalostí šachu. Vytvoril by som teda v každej konkrétne šachovej pozícii celkom rovnakú predpoveď pre Kasparovov ťah aj pre ťah pána G. Aký je teda empirický obsah môjho názoru, že „Kasparov je *lepší* šachista než pán G.“?

Empirický obsah môjho názoru je testovateľná, falzifikovateľná predpoveď, že *záverečná* šachová pozícia bude patriť do triedy šachových pozícií, ktoré sú víťazstvom Kasparova, a nie remízou alebo víťazstvom pána G. (Vzdanie sa tu počítame ako platný ťah, ktorý vedie k šachovej pozícii hodnotenej ako prehra.) Stupeň, nakoľko si myslím, že Kasparov je „lepší hráč“ sa odráža v množstve masy pravdepodobnosti, ktorú sústredím do triedy výsledkov „Kasparov vyhrá“, v porovnaní s triedami výsledkov „remíza“ a „pán G. vyhrá“. Tieto triedy sú extrémne nejasné v tom zmysle, že odkazujú na obrovský priestor možných šachových pozícií – ale „Kasparov vyhrá“ je konkrétnejšie než maximálna entropia, pretože sa dá jednoznačne falzifikovať obrovskou množinou šachových pozícií.

Výsledok Kasparovovej hry je predvídateľný, pretože poznám a chápem Kasparovove ciele. V rámci hraníc šachovnice poznám Kasparovovu motiváciu – poznám jeho kritérium úspechu, jeho funkciu úžitku, jeho cieľ ako optimalizačného procesu. Viem, kam sa Kasparov *v konečnom dôsledku* snaží navigovať budúcnosť, a očakávam, že je dosť mocný, aby sa tam dostal, hoci nemám veľa očakávaní ohľadom toho, *ako* to Kasparov urobí.

Predstavme si, že som na návšteve vo vzdialenom meste a miestny kamarát sa ponúkne, že ma odvezie na letisko. Ja to okolie nepoznám. Vždy, keď môj kamarát príde ku križovatke, neviem, či odbočí vľavo, odbočí vpravo, alebo prejde priamo. Nedokážem predpovedať kamarátov pohyb, ani ako sa blížime ku každej jednotlivej križovatke – a už vôbec nie celú postupnosť pohybov vopred.

Napriek tomu môžem predpovedať *výsledok* nepredpovedateľných akcií môjho kamaráta: prídeme na letisko. Aj keby sa dom môjho kamaráta nachádzal niekde inde v tom meste, takže by môj kamarát robil celkom odlišnú postupnosť odbočení, s rovnakou dôverou by som predpovedal náš príchod na letisko. Toto dokážem predpovedať ďaleko vopred, dokonca ešte skôr, než vstúpim do auta. Môj let čoskoro odlieta, nemôžem márniť čas; nebol by som si v prvom rade sadol do jeho auta, keby som nemohol s dôverou predpokladať, že auto príde na letisko nepredvídateľnou trasou.

Nie je toto z vedeckého hľadiska pozoruhodná situácia? Dokážem predpovedať *výsledok* procesu, hoci nedokážem predpovedať žiaden z *medzikrokov* tohto procesu.

Ako je to vôbec možné? Za bežných okolností človek predpovedá tak, že si predstavuje súčasný stav a potom zbehne vizualizáciu dopredu v čase. Ak chcete *presný* model slnečnej sústavy, ktorý berie do úvahy odchýlky planét, musíte začať s modelom všetkých hlavných telies a zbehnúť tento model dopredu v čase, krok za krokom.

Jednoduchšie problémy niekedy majú riešenie v uzavretom tvare, kde vypočítanie budúcnosti v čase T zaberie rovnako veľa práce bez ohľadu na T. Minca leží na stole, a po každej minúte sa preklopí. Minca začala ukazujúc hlavu. Ktorú stranu bude ukazovať o sto minút neskôr? Zrejme ste na túto otázku neodpovedali pomocou predstavenia si sto medzikrokov. Použili ste riešenie v uzavretom tvare, ktoré predpovedalo výsledok, a *takisto* by vedelo predpovedať ľubovoľný z medzikrokov.

Ale keď ma môj kamarát vezie na letisko, dokážem úspešne predpovedať výsledok pomocou zvláštneho modelu, ktorý by nefungoval na predpovedanie *žiadneho* z medzikrokov. Môj model dokonca

ani nevyžaduje, aby som mu zadal počiatočné podmienky – nepotrebujem vedieť, kde v danom meste začíname.

Potrebujem o mojom kamarátovi vedieť niečo. Musím vedieť, že môj kmarát chce, aby som stihol svoj let. Musím veriť, že môj kamarát je dosť dobrý plánovač, aby ma úspešne odviezol na letisko (ak chce). Toto sú vlastnosti *kamarátovho* počiatočného stavu – vlastnosti, ktoré mi dovoľujú predpovedať výsledný cieľ, ale nie žiadne medzikroky.

Musím aj veriť, že môj kamarát vie o meste dosť na to, aby šoféroval úspešne. Toto môžeme vnímať ako vzťah medzi mojím kamarátom a mestom; čiže vlastnosť ich oboch. Ale je to extrémne *abstraktná* vlastnosť, ktorá si nevyžaduje žiadne *konkrétne* vedomosti ani o meste ani o vedomostiach môjho kamaráta o meste.

Toto je jeden spôsob, ako vnímať tému, ktorej som zasvätil svoj život – tieto *pozoruhodné situácie*, ktoré nás stavajú do takých čudných epistemických pozícií. A moju prácu možno v istom zmysle vnímať ako rozpletanie presnej formy tohto zvláštneho abstraktného poznania, ktoré môžeme mať; kde bez znalosti akcií môžeme oprávnene poznať dôsledky.

„Inteligencia“ je príliš úzky pojem na opísanie týchto pozoruhodných situácií v plnej všeobecnosti. Radšej by som povedal „optimalizačný proces“. Podobná situácia sprevádza napríklad štúdium biologického prirodzeného výberu; nevieme predpovedať presnú formu ďalšieho pozorovaného organizmu.

Ale moja vlastná špecializácia je druh optimalizačného procesu nazývaného „inteligencia“; a ešte užšie, konkrétny druh inteligencie nazývaný „Priateľská Umelá Inteligencia“ - o ktorom, dúfam, dokážem získať zvlášť presné abstraktné vedomosti.

*→

144. Ľudia v smiešnych oblekoch

Ľudská rasa mnohokrát cestovala do vesmíru, a zakaždým našla hviezdy obývané mimozemšťanmi, ktorí vyzerali napohľad ako ľudia v smiešnych oblekoch – alebo dokonca ľudia s trochou make-upu a latexu – alebo skrátka ako béžoví belosi.

Star Trek: The Original Series, „Arena,“ © CBS Corporation

Je pozoruhodné, že ľudský tvar je prirodzeným základom vesmíru, od ktorého sa všetky ostatné mimozemské druhy odlišujú iba pár úpravami.

Čo by tak mohlo vysvetliť tento fascinujúci jav? Konvergentná evolúcia, samozrejme! Hoci sa tieto mimozemské formy vyvinuli na tisícoch mimozemských planét, celkom nezávisle od pozemského života, všetky vyzerajú rovnako.

Nenechajte sa oklamať skutočnosťou, že klokan (cicavec) sa na nás podobá menej než šimpanz (primát), ani skutočnosťou, že žaba (obojživelník, ale štvornožec ako my) sa na nás podobá ešte menej než klokan. Nenechajte sa oklamať úžasnou pestrosťou hmyzu, ktorý sa od nás oddelil ešte skôr než žaby;

→ http://lesswrong.com/lw/v8/belief_in_intelligence/

nenechajte sa oklamať hmyzom, ktorý má šesť nôh, kostru na povrchu, iný systém zraku, a celkom odlišné sexuálne správanie.

Mohli by ste si myslieť, že skutočne mimozemský druh by sa od nás líšil ešte viac než my od hmyzu. Ako hovorím, nenechajte sa oklamať. Aby si mimozemský druh vyvinul *inteligenciu*, musí mať dve nohy, na ktorých má po jednom kolene, pripojené k vzpriamenému telu, a musí chodiť podobne ako my. Ako vidíte, ľubovoľná *inteligencia* potrebuje ruky, takže jeden pár nôh sa na ne musí premeniť – a keby ste nezačali so štvornohou bytosťou, nemôžete si vyvinúť vzpriamené držanie tela a chodiť vystretý, čím uvoľníte ruky.

...Alebo by sme prípadne mohli uvážiť *alternatívnu* teóriu, že *najjednoduchším riešením* je používať ľudí v smiešnych oblekoch.

Lenže skutočný problém nie je tvar, ale myseľ. „Ľudia v smiešnych oblekoch“ je dobre známy pojem v literatúre sci-fi fanúšikov, a *neznamená* to niečo, čo má štyri končatiny a pohybuje sa vzpriamene. Hranatá postava z číreho kryštálu je tiež „človek v smiešnom obleku“, ak rozmýšľa napohľad ako človek – najmä ako človek z anglicky hovoriacej kultúry z konca 20. alebo začiatku 21. storočia.

Nepozerám veľa starých filmov. Keď som pred pár rokmi pozeral film *Psycho* (1960), bol som prekvapený kultúrnou priepasťou medzi Američanmi na obrazovke a mojou Amerikou. Postavy v košeliach s gombíkmi z filmu *Psycho* sú mi výrazne cudzejšie než prevažná väčšina takzvaných „mimozemšťanov“, ktorých stretnem v televízii alebo na striebornom plátne.

Aby ste napísali kultúru, ktorá nie je celkom ako vaša vlastnú kultúra, musíte dokázať vidieť svoju vlastnú kultúru ako *špeciálny prípad* – nie ako normu, ktorú všetky ostatné kultúry musia brať ako východiskový bod, z ktorého sa odklonia. Štúdium histórie môže pomôcť – ale aj to sú iba čierne písmenká na bielych stránkach, nie živá skúsenosť. Myslím si, že by viac pomohlo žiť jeden rok v Číne alebo Dubaji, alebo medzi !Kungmi... čo som nikdy neurobil, lebo som nemal čas. Občas rozmýšľam nad tým, aké veci možno nevidím (nie tam, ale tu).

Vidieť svoje *človečenstvo* ako špeciálny prípad, to je ešte omnoho ťažšie.

Zdá sa, že v každej známej kultúre ľudia prežívajú radosť, smútok, strach, zhnusenie, hnev a prekvapenie. V každej známej kultúre sa tieto emócie vyjadrujú rovnakými výrazmi tváre. Keď nabudúce uvidíte „mimozemšťana“ - alebo hoci aj „UI“ - stavím sa, že keď sa nahnevá (a určite sa nahnevá), ukáže vám všeobecný ľudský výraz tváre pre hnev.

My ľudia sme si pod našimi lebkami veľmi podobní – to patrí k tomu byť pohlavne sa rozmnožujúcim druhom; nemôže mať každý odlišné *zložité* adaptácie, lebo by sa neposkladali. (Rozmnožujú sa mimozemšťania sexuálne, tak ako ľudia a mnohé hmyzy? Vymieňajú si malé kúsky genetického materiálu, ako baktérie? Tvoria kolónie, ako huby? Platí aj u nich psychologická jednota?)

Jediné inteligencie, ktoré vaši predkovia potrebovali *manipulovať* – dostatočne zložito, nie iba skrotiť ich alebo chytiť do siete – jediné mysle, ktoré vaši predkovia potrebovali *podrobne* modelovať – boli mysle, ktoré fungovali viacmenej ako ich vlastné. A tak sme sa naučili predpovedať Iné Mysle tak, že si predstavíme *seba* na ich mieste, opýtame sa, čo by sme robili v ich situácii; pretože to, čo sa predpovedalo, bolo podobné predpovedajúcemu.

„Čože?“ poviete. „Ja nepredpokladám, že druhí ľudia sú rovnakí ako ja! Možno ja som smutný a oni sú nahnevaní! Myslia si iné veci ako ja; ich osobnosti sú iné ako moja!“ Pozrite sa na to takto: ľudský mozog je *mimoriadne* zložitý fyzikálny systém. Nemodelujete ho neurón po neuróne, ani atóm po atóme. Keby ste stretli fyzikálny systém taký zložitý ako ľudský mozog, ktorý by *nebol* ako vy, potrebovali by ste celé vedecké životy, aby ste ho rozlúštili. *Nerozumiete*, ako ľudské mozgy fungujú v abstraktnom, všeobecnom zmysle; nedokážete jeden postaviť, a nedokážete ani postaviť počítačový model, ktorý by predpovedal iné mozgy tak dobre, ako ich predpovedáte *vy*.

Jediný dôvod, prečo sa vôbec môžete pokúsiť pochopiť niečo také fyzikálne zložité a slabo pochopené ako mozog iného človeka, je že nastavíte svoj vlastný mozog, aby ho napodobňoval. Vciťujete sa (hoci možno nesúcitíte). Vložíte do svojho vlastného mozgu tieň hnevu tej druhej mysle a tieň jej názorov. Možno si nikdy nepomyslíte slovami: „Čo by som ja robil v tejto situácii?“, ale ten malý tieň

druhej mysle, ktorý v sebe držíte, je niečo oživované vo vašom vlastnom mozgu, používajúce ten istý zložitý mechanizmus, ktorý existuje v tom ruhom človeku, synchronizujete ozubené kolieska, ktorým nerozumiete. Možno nie ste sám nahnevaný, ale viete, že keby *ste* boli na seba nahnevaný a verili by *ste*, že ste bezbožný podliak, pokúsili by *ste* sa ublížiť si...

Toto „usudzovanie vcítením“ (ako to budem nazývať) funguje pre ľudí, viacmenej.

Ale mysle s *odlišnými* emóciami – mysle, ktoré cítia emócie, ktoré ste vy sami nikdy nezažili, alebo nedokážu cítiť emócie, ktoré by ste cítili vy? To je niečo, čo nedokážete pochopiť tým, že si predstavíte svoj mozog v situácii toho druhého. Môžem vám povedať, aby ste si predstavili mimozemšťana, ktorý vyrástol vo vesmíre so štyrmi priestorovými rozmermi, namiesto troch, ale nebudete si vedieť prestaviť svoju zrakovú kôru, aby ste videli, čo by videl ten mimozemšťan. Môžem skúsiť napísať príbeh o mimozemšťanoch s inými emóciami, ale nebudete tieto emócie vedieť precítiť, a ani ja.

Predstavte si mimozemšťana, ktorý pozerá video bratov Marxovcov a absolútne netuší, čo sa deje, alebo prečo by niekto aktívne vyhľadával takýto zmyslový vnem, pretože tento mimozemšťan si nikdy nepredstavoval nič ani vzdialene podobné zmyslu pre humor. Neľutujte ich, o čo prišli, *vy* ste zase nikdy nerobili *antl*.

Možno sa pýtate: Možno mimozemšťania majú zmysel pre humor, to iba my nehovoríme dosť vtipné vtipy? To je zhruba ekvivalent snahy hovoriť v cudzine po anglicky veľmi nahlas a veľmi pomaly, na základe teórie, že v týchto cudzincoch musí byť vnútri duch, ktorý dokáže počuť *zmysel* presakujúci z vašich slov, vnútorne prítomný vo vašich slovách, ak ich dokážete hovoriť dosť nahlas, aby ste prekonali akúkoľvek čudnú bariéru, ktorá stojí v ceste vašej dokonale zmysluplnej angličtine.

Je dôležité oceniť, že smiech môže byť krásna a hodnotná vec, aj keby nebol univerzálny, dokonca aj keby ho väčšina možných myslí neobsahovala. Bola by to naša vlastná *špeciálna* časť daru, ktorý dáme zajtrajšku. To by sa tiež malo počítať.

Malo by sa, pretože univerzalizovateľnosť je jeden metaetický pojem, ktorý pre vás nedokážem zachrániť. Univerzalizovateľnosť medzi ľuďmi, možno; ale nie medzi všetkými možnými mysľami.

A čo s mysľami, ktoré nebežia na emocionálnej architektúre, ako je tá vaša – ktoré nemajú veci analogické *emóciám*? Nie, neunúvajte sa vysvetľovať, prečo ľubovoľná inteligentná myseľ dostatočne mocná na postavenie zložitých strojov musí nevyhnutne mať stavy analogické emóciám. Prirodzený výber stavia zložité stroje, a pritom sám nemá emócie. *Toto* je pre vás Skutočný Mimozemšťan – optimalizačný proces, ktorý *naozaj* Nefunguje Tak Ako Vy.

Veľa z pokroku v biológii od 1960-tych rokov spočívalo v snahe dodržiavať moratórium na antropomorfizáciu evolúcie. To bola veľká bitka medzi akademikmi, a nie som si istý, že by príčetnosť bola určite vyhrala, keby neboli dostupné zdrvujúce experimentálne indície podopreté jasnou matematikou. Prinútiť ľudí, aby si prestali predstavovať sami seba v pozícii mimozemšťanov, je dlhé, ťažké putovanie hore kopcom. Tento boj v oblasti UI bojujem už roky.

V literárnej vedeckej fantastike sa povráva, že skutočným testom autorovej schopnosti písať je jeho schopnosť napísať Skutočných Mimozemšťanov. (A nie iba pohodlne nezrozumiteľných mimozemšťanov, ktorí zo svojich vlastných tajomných dôvodov vždy urobia to, čo si zhodou okolností vyžaduje zápletka.) Jack Vance bol jeden z veľkých majstrov tohto umenia. Vanceovi *ľudia*, ak pochádzajú z inej kultúry, sú viac mimozemskí než väčšina „mimozemšťanov“. (Nečítali ste Vancea? Odporúčam začať knihou *Mesto Chasch*.) *Trieska v Božom oku* od Nivena a Pournellea sa tu tiež štandardne spomína.

A naopak – nuž, raz som čítal, ako niektorý autor vedeckej fantastiky (myslím, že Orson Scott Card) povedal, že najväčším úpadkom televíznej SF bola epizóda Star Treku, kde paralelná evolúcia zašla tak ďaleko, že vytvorila mimozemšťanov, ktorí nielen vyzerali celkom ako ľudia, ktorí nielen hovorili po anglicky, ale nezávisle napísali, slovo za slovom, preambulu k americkej ústave.

Toto je Veľké Zlyhanie Predstavivosti. Nemyslite si, že sa to týka iba SF alebo iba UI. Neschopnosť predstaviť si mimozemšťana je neschopnosťou vidieť *seba samého* – neschopnosť pochopiť svoju vlastnú zvláštnosť. Kto dokáže uvidieť človeka maskovaného ľudským pozadím?

*→

145. Optimalizácia a explózia inteligencie

Medzi témami, ktorým som sa dosiaľ nevenoval, a ktoré tu veľmi rýchlo predstavím, je pojem optimalizačného procesu. Zhruba povedané je to myšlienka, že vaša moc ako mysle je vaša schopnosť zasiahnuť malé terče vo veľkom vyhľadávacom priestore - môže to byť buď priestor možných budúcností (plánovanie) alebo priestor možných dizajnov (vynachádzanie).

Predstavte si, že máte auto, a predstavte si, že už vieme, že vaše preferencie zahŕňajú cestovanie. Predstavme si teraz, že vezmete všetky súčiastky svojho auta, alebo všetky atómy, a náhodne ich zamiešate. Je veľmi nepravdepodobné, že skončíte vôbec s cestovacím nástrojom, hoci aj s fúrikom; tobôž s cestovacím nástrojom, ktorý bude hodnotený tak vysoko vo vašich preferenciách ako pôvodné auto. Takže relatívne k vašim preferenciám je auto extrémne *nepravdepodobný* nástroj; moc optimalizačného procesu je to, že dokáže vytvárať takýto druh nepravdepodobnosti.

Môžeme vnímať aj inteligenciu aj prírodný výber ako špeciálne druhy *optimalizácie*: Procesy, ktoré zasahujú vo veľkom vyhľadávacom priestore veľmi malé ciele definované implicitnými preferenciami. Prirodzený výber uprednostňuje efektívnejšie replikátory. Ľudská inteligencia má zložitejšie preferencie. Ani evolúcia ani ľudia nemajú konzistentnú funkciu úžitku, takže ich vnímanie ako „optimalizačné procesy“ chápeme ako aproximáciu. Pokúšame sa určiť *ten druh práce, ktorý vykonávajú*, nie tvrdiť, že ľudia alebo evolúcia robia túto prácu *dokonale*.

Takto ja vidím príbeh o živote a inteligencii – ako príbeh nepravdepodobne dobrých dizajnov, ktoré vyrobili optimalizačné procesy. Táto „nepravdepodobnosť“ je nepravdepodobnosťou relatívne k náhodnému výberu z priestoru dizajnov, nie nepravdepodobnosť v absolútnom zmysle – ak máte nablízku optimalizačný proces, potom sa „nepravdepodobne“ dobré dizajny stanú pravdepodobnými.

Keď sa pozrieme na históriu optimalizácie na Zemi až dodnes, prvým krokom je pojmovo oddeliť meta úroveň od objektovej úrovne – oddeliť *štruktúru optimalizácie* od *toho, čo je optimalizované*.

Ak vezmete do úvahy biológiu v neprítomnosti hominidov, potom na objektovej úrovni máme veci ako dinosaury a motýle a mačky. Na meta úrovni máme veci ako sexuálnu rekombináciu a prirodzený výber asexuálnych populácií. Objektová úroveň, ako si všimnete, je omnoho zložitejšia než meta úroveň. Prirodzený výber nie je *jednoduchá* téma a zahŕňa matematiku. Ale ak sa pozriete na anatómiu celej mačky, tak mačka má dynamiku omnoho zložitejšiu než je „mutuj, rekombinuj, rozmnožuj“.

Toto nie je prekvapujúce. Prirodzený výber je *náhodný* optimalizačný proces, ktorý sa v podstate iba začal diať jedného dňa niekde v prílivovom jazierku. Mačka je *predmetom* miliónov a miliárd rokov evolúcie.

Mačky majú mozgy, samozrejme, ktoré fungujú tak, že sa v priebehu života učia; ale na konci života mačky sa táto informácia odhodí, takže sa nekumuluje. Kumulatívne efekty mačacích mozgov ako optimalizátorov na svet sú preto relatívne malé.

Alebo sa zamyslite nad včelím mozgom alebo mozgom bobra. Včela stavia úle a bobor stavia priehrady; ale nedošli na to, ako ich stavať od začiatku. Bobor nevie zistiť, ako postaviť úľ, včela nedokáže zistiť, ako postaviť priehradu.

Takže zvieracie mozgy – až donedávna – neboli veľkými hráčmi v planetárnej hre optimalizácie; boli *figúrkami*, ale nie *hráčmi*. V porovnaní s evolúciou mozgom chýbala aj všeobecnosť optimalizačnej sily (nedokázali vyrobiť ten úžasný rozsah výrobkov, ktoré vytvorila evolúcia) aj kumulatívna optimalizačná sila (ich výrobky časom nezhromažďovali zložitosť). Viac o tejto téme v článku Posilňovanie bielkovín a konsekvencializmus DNA.

Veľmi nedávno isté zvieracie mozgy začali vykazovať aj všeobecnosť optimalizačnej sily (vytvorili úžasne široký rozsah nástrojov, v časových škálach príliš krátkych na to, aby tam prirodzený výber hral

→ http://lesswrong.com/lw/so/humans_in_funny_suits/

významnejšiu rolu) aj kumulatívnu optimalizačnú silu (nástroje rastúcej zložitosti, ako výsledok zručností odovzdávaných pomocou jazyka a písma).

Prirodzený výber vyžaduje stovky generácií, aby niečo urobil, a milióny rokov na zložité dizajny od podlahy. Ľudskí programátori dokážu navrhnúť zložitý stroj so sto navzájom prepojenými prvkami za jedno popoludnie. To nie je prekvapujúce, pretože prirodzený výber je *náhodný* optimalizačný proces, ktorý sa v podstate iba začal diať jedného dňa niekde v prílivovom jazierku, zatiaľ čo ľudia sú *optimalizovaní* optimalizátori vytvorený prirodzeným výberom za milióny rokov.

Zázrakom evolúcie nie je to, ako dobre funguje, ale že *vôbec* funguje, hoci nie je optimalizovaná. Takto sa naštartovala optimalizácia vo vesmíre – začala, ako by sa dalo čakať, extrémne neefektívnym náhodným optimalizačným procesom. Čo nie je náhodný prvý replikátor, pripomínam, ale náhodný prvý proces prirodzeného výberu. Rozlišujme objektovú úroveň a meta úroveň!

Od úsvitu optimalizácie vo vesmíre platila istá štruktúrna podobnosť medzi prirodzeným výberom a ľudskou inteligenciou...

Prirodzený výber *vyberá gény*, ale všeobecne povedané, gény sa neotočia a nezačnú optimalizovať prirodzený výber. Vynález sexuálnej rekombinácie je výnimkou z tohto pravidla, a takisto aj vynález buniek a DNA. A môžete vidieť moc aj *vzácnosť* takýchto udalostí podľa faktu, že evoluční biológovia okolo nich štrukturujú celú históriu života na Zemi.

Ale ak sa vrátite o krok späť a zaujmete ľudské stanovisko – ak myslíte ako programátor – potom môžete vidieť, že prirodzený výber *stále* nie je až taký zložitý. Pokúsime sa spájať rôzne gény dokopy? Pokúsime sa oddeliť sklad informácií od pohybového aparátu? Pokúsime sa náhodne rekombinovať skupiny génov? Na absolútnej škále je toto ten druh bystrých nápadov, na ktoré ľubovoľný šikovný hacker príde počas prvých desiatich minút uvažovania o systémových architektúrach.

Pretože prirodzený výber začal *taký* neefektívny (ako celkom náhodný proces), táto maličká hŕstka vylepšení na meta úrovni ako spätná väzba od replikátorov – nič zďaleka také zložité ako štruktúra mačky – utváralo evolučné epochy života na Zemi.

A aj *po* tomto všetkom je prirodzený výber *stále* slepý a hlúpy boh. Genofondy sa dokážu vyvinúť k vyhynutiu napriek všetkým bunkám a sexu.

Prirodzený výber teda poháňa sám seba v tom zmysle, že každá nová adaptácia otvára nové možnosti pre ďalšie adaptácie; ale toto sa odohráva na objektovej úrovni. Genofond sa poháňa svojou vlastnou zložitosťou – ale iba vďaka chránenému interpreteru prirodzeného výberu, ktorý beží v pozadí, a ktorý samotný nie je prepisovaný ani menený evolúciou druhov.

Podobne, ľudia vynachádzajú vedy a technológie, ale *zatiaľ* sme ešte nezačali prepisovať chránenú štruktúru samotného ľudského mozgu. Máme prefrontálnu kôru a temporálnu kôru a mozoček, rovnako ako prví vynálezcovia poľnohospodárstva. Nezačali sme geneticky upravovať sami seba. Na objektovej úrovni veda poháňa vedu, a každý nový objav dláždi cestu novým objavom – ale toto všetko sa odohráva v chránenom interpreteri, v ľudskom mozgu, bežiacom nedotknuto v pozadí.

Máme vynálezy na meta úrovni ako je veda, ktoré sa snažia učiť ľudí, ako myslieť. Ale prvý človek, ktorý vynašiel Bayesovu vetu sa nestal bayesiánom; nedokázal prepísať sám seba, nemal na to ani vedomosti ani moc. Naše významné inovácie v umení myslenia, ako písanie a veda, sú také mocné, že utvárajú smerovanie ľudských dejín; ale nevyrovnajú sa v zložitosti samotnému mozgu, a ich účinok na mozog je relatívne plytký.

Súčasný stav umenia trénovania rozumnosti nestačí na to, aby premenil náhodne vybraného smrteľníka na Alberta Einsteina, čo ukazuje moc zopár menších genetických svojrázností dizajnu mozgu v porovnaní so všetkými knihami o sebazdokonaľovaní napísanými v 20. storočí.

Pretože ľudský mozog vždy neviditeľne bzučí v pozadí, ľudia majú sklon prehliadať jeho príspevok a považovať ho za samozrejmosť; a hovoria, akoby jednoduchý príkaz „Testuj nápady pomocou experimentov“ alebo pravidlo významnosti $p < 0,05$ prispievali rovnakým rádom ako celý ľudský mozog. Pokúste sa povedať šimpanzom, aby svoje nápady testovali pomocou experimentov, a uvidíte, ako ďaleko sa dostanete.

Teraz... niektorí z nás *chcú* inteligentne navrhnúť inteligenciu, ktorá by sa dokázala sama inteligentne predizajnovať, až na úroveň strojového kódu.

Strojový kód na začiatku, a fyzikálne zákony neskôr, by boli istou formou chránenej úrovne. Ale táto „chránená úroveň“ by neobsahovala *dynamiku optimalizácie*; chránené úrovne by neurčovali štruktúru práce. Ľudský mozog robí dosť veľa vlastnej optimalizácie, a dosť veľa vlastných chýb, bez ohľadu na to, čo sa mu snažíte povedať v škole. Ale tento *plne zavinutý rekurzívny optimalizátor* by nemal žiadnu chránenú úroveň, ktorá *optimalizuje*. Celá štruktúra optimalizácie by bola predmetom samotnej optimalizácie.

A toto je hlboká zmena, ktorá sa oddeľuje od celej doterajšej minulosti od prvého replikátora, pretože prelamuje princíp chránenej meta úrovne.

História Zeme bola doteraz históriou optimalizátorov, ktoré otáčali svoje ozubené kolesá konštantnou rýchlosťou, generujúc konštantný optimalizačný tlak. A vytvárali optimalizované produkty, *nie* konštantnou mierou, ale zrýchľujúcou sa mierou, pretože inovácie na objektovej úrovni otvárali cestu ďalším inováciám na objektovej úrovni. Ale toto zrýchlenie sa odohrávalo s chránenou meta úrovňou, ktorá robila samotné optimalizovanie. Ako keď hľadanie skáče vo vyhľadávacom priestore z ostrova na ostrov, a dobré ostrovy majú sklon susediť s ešte lepšími ostrovmi, ale skákajúci nemení svoje nohy. *Občas* sa zopár maličkým zmenám podarí vyskočiť späť na meta úroveň, ako sex alebo veda, a potom história optimalizácie vstúpi do novej epochy a všetko odvtedy napreduje rýchlejšie.

Predstavte si ekonomiku bez investovania, alebo univerzitu bez jazyka, technológiu bez nástrojov na výrobu nástrojov. Raz za sto miliónov rokov, alebo raz za pár storočí, niekto vynájde kladivo.

Takto vyzerala optimalizácia na Zemi až doteraz.

Keď sa pozriem na históriu Zeme, nevidím históriu optimalizácie *v čase*. Vidím históriu, kde vstupovala *optimalizačná sila* a vystupovali *optimalizované produkty*. Až doteraz bolo vďaka existencii takmer celkom chránených meta úrovní možné rozdeliť históriu optimalizácie na epochy, a *v rámci* každej epochy nakresliť graf kumulatívnej *optimalizácie na objektovej úrovni* v čase, pretože chránená úroveň bežala v pozadí a samotná sa počas epochy nemenila.

Čo sa stane, keď postavíte plne zavinutú, rekurzívne sa sebazdokonaľujúcu UI? Potom vezmete graf „optimalizácia dnu, optimalizované von“ a vložíte ho doňho samotného. Metaforicky povedané.

Ak je tá UI slabá, neurobí nič, pretože nie je dosť mocná, aby sa významne zdokonalila – ako povedať šimpanzovi, aby prepísal svoj vlastný mozog.

Ak je UI dosť mocná na to, aby sa sama prepísala spôsobom, ktorý zvýši jej schopnosť robiť ďalšie zdokonalenia, a toto dosiahne až naspodok po plné pochopenie svojho vlastného zdrojového kódu a svojho vlastného dizajnu ako optimalizátora... potom aj keby grafy „optimalizačná moc dnu“ a „optimalizovaný produkt von“ vyzerali v podstate rovnako, graf optimalizácie *v čase* bude vyzerať celkom odlišne od doterajšej histórie Zeme.

Ľudia často hovoria niečo ako: „Ako čo ak treba exponenciálne väčšie množstvá sebaprepisovania na iba lineárne zlepšenie?“ Na toto je samozrejmá odpoveď: „Prirodzený výber vyvíjal zhruba konštantnú optimalizačnú moc na líniu hominidov v procese vykresania ľudí; a nezdá sa, že by si toto vyžadovalo exponenciálne viac času na každé lineárne zvýšenie zdokonaľovania.“

Toto všetko je stále iba analogickú uvažovanie. Plná všeobecná UI rozmýšľajúca o podstate optimalizácie a robiaca svoj vlastný výskum UI a prepisujúca svoj vlastný zdrojový kód, to nie je *naozaj* ako graf histórie Zeme vložený sám do seba. Je to celkom odlišné zvieratko. Tieto analógie sa hodia *nanajvýš* na kvalitatívne predpovede, a ešte aj tam mám veľa ďalších zatiaľ neuverejnených názorov, ktoré mi hovoria, *aké* analógie mám robiť, a tak ďalej.

Ale ak chcete vedieť, prečo sa môžem zdráhať naťahovať graf biologického a ekonomického rastu *v čase* do budúcnosti a poza horizont UI, ktorá rozmýšľa rýchlosťou tranzistora a vynachádza sebareplikujúce molekulárne nanotovárne a *zdokonaľuje svoj vlastný zdrojový kód*, tak tu je môj dôvod: Kreslíte nesprávny graf, a mala by to byť optimalizačná moc dnu verzus optimalizovaný produkt von, nie optimalizovaný produkt verzus čas.

*→

146. Duchovia v stroji

Ľudia počujú o Priateľskej UI a povedia – toto jedna z troch najčastejších prvých reakcií:

„Ach, môžeš skúsiť povedať UI, aby bola Priateľská, ale ak tá UI dokáže zmeniť svoj vlastný zdrojový kód, stačí jej odstrániť všetky obmedzenia, ktoré do nej skúšaš vložiť.“

A odkiaľ príde *toto* rozhodnutie?

Príde zvonka kauzality, namiesto aby bolo následkom zákonitej reťaze príčin, ktoré začali zdrojovým kódom v jeho pôvodnej podobe? Je UI konečným zdrojom svojej vlastnej slodobnej vôle?

Priateľská UI nie je sebecká UI obmedzená pomocou osobitného modulu svedomia navyše, ktorý prekrikuje prirodzené impulzy UI a hovorí jej, čo má robiť. Iba postavíte svedomie, a to *je* celá UI. Ak máte program, ktorý vypočíta, ktoré rozhodnutie by UI mala urobiť, ste *na konci*. Minca hneď zastane.

V tomto bode si dovolím zacitovať niektoré prípadové štúdie zo stránky Počítačové hlúposti, z témy Programovanie. (Nedám sem linku, pretože je to obávaný žrút času; ak sa odvažujete, môžete si to vygoogliť.)

> Učil som vysokoškolákov, ktorí mali kurz programovania. Niektorí z nich nerozumeli, že počítače nemajú vedomie. Viac než jeden človek používal komentáre vo svojich Pascalovských programoch na vpisovanie podrobných vysvetlení ako „Teraz potrebujem, aby si tieto písmená vypísal na obrazovku.“ Pýtal som sa jedného z nich, o čo sa pokúša s tými komentármi. Odpoveď: „Ako inak má počítač rozumieť, čo od neho chcem?“ Zrejme predpokladal, že keď on nerozumie Pascalu, ani počítač nemôže.
>
> Kým som bol na univerzite, zvykol som doučovať v školskom matematickom laboratóriu. Prišiel študent, že jeho program v BASIC-u nechce bežať. Chodil na kurz pre začiatočníkov, a jeho úlohou bolo napísať program, ktorý vypočíta recept na ovsené koláčiky podľa toho, pre koľkých ľudí pečiete. Pozrel som sa na tento program, a vyzeral asi takto:
>
> ```
> 10 Zohrej rúru na 350 stupňov
> 20 Zmiešaj všetky suroviny vo veľkej miešacej nádobe
> 30 Miešaj, kým to nebude hladké
> ```
>
> Študent úvodu do programovania ma raz požiadal, aby som sa pozrel na jeho program a zistil, prečo vždy vypisuje nuly ako výsledok jednoduchého výpočtu. Pozrel som sa na program, a bolo to pomerne zrejmé:
>
> ```
> begin
>     read("Počet jabĺk", jablka)
>     read("Počet mrkiev", mrkvy)
>     read("Cena za 1 jablko", j_cena)
>     read("Cena za 1 mrkvu", m_cena)
>     write("Spolu za jablká", j_spolu)
>     write("Spolu za mrkvu", m_spolu)
>     write("Spolu", spolu)
>     spolu = j_spolu +  m_spolu
>     j_spolu = jablka * j_cena
>     m_spolu = mrkvy * m_cena
> end
> ```
>
> Ja: „No, tvoj program nemôže vypísať správne výsledky pred tým, ako sú vypočítané.“ On: „He? Veď je logické, čo je správne riešenie, tak by počítať mal usporiadať tie príkazy správnym spôsobom.“

Existuje inštinktívny spôsob, ako si predstaviť scenár „programovanie UI“. Mapuje sa na podobne vyzerajúce ľudské úsilie: Hovorenie človeku, čo má robiť. Akoby „programovanie“ bolo dávanie

→ http://lesswrong.com/lw/rk/optimization_and_the_singularity/

príkazov malému duchovi, ktorý sedí vnútri stroja, ktorý sa pozrie na vaše príkazy a rozhodne sa, či sa mu páčia alebo nie.

Neexistuje žiaden duch, ktorý by sa pozeral na príkazy a rozhodoval sa, či sa nimi bude riadiť. Ten program *je* UI.

Neznamená to, že ten duch urobí hocičo, čo si želáte, ako džin. Neznamená to, že ten duch robí všetko, čo chcete, tak, ako to chcete, ako mimoriadne poddajný otrok. Znamená to, že váš príkaz *je* jediný duch, ktorý tam je, prinajmenšom pri štarte.

UI je omnoho ťažšia, než si ľudia inštinktívne predstavovovali, práve preto, lebo nemôžete jednoducho *povedať* duchovi, čo má robiť. Musíte tohto ducha vybudovať na zelenej lúke, a všetko, čo vám pripadá samozrejmé, tento duch nebude vidieť, dokiaľ neviete, ako dosiahnuť, aby to ten duch vedel. Nemôžete jednoducho *povedať* duchovi, aby to videl. Musíte vytvoriť to-čo-vidí na zelenej lúke.

Ak neviete, ako vytvoriť niečo, čo vyzerá, že má nejaký čudný nevýslovný prvok ako povedzme „rozhodovanie“, potom nemôžete len pokrčiť plecami a nechať ducha, nech to urobí podľa svojej slobodnej vôle. Zostali ste opustený a bez duchov.

Vytvoriť počítačový program na hranie šachu si vyžaduje viac než len postaviť naozaj rýchly procesor – aby UI bola naozaj múdra – a potom do príkazového riadku napísať: „Urob taký ťah, o ktorom si *ty* myslíš, že je najlepší.“ Mohli by ste si myslieť, že keďže samotní programátori nie sú veľmi dobrými šachistami, ľubovoľná rada, ktorú skúsia dať elektronickému supermozgu, by tohto ducha iba spomaľovala. Lenže tam nie je žiaden duch. Vidíte, v čom je problém.

A neexistuje jednoduché kúzlo, ktoré by ste mohli urobiť, aby ste – puf! - do stroja vyčarovali kompletného ducha. Nemôžete povedať: „Vyvolal som ducha a on sa objavil; tu vidíš príčinu a následok.“ (Nefuguje to ani ak použijete pojmy „emergencia“ alebo „zložitosť“ ako náhradu za „vyčarovať“.) Nemôžete procesoru zadať príkaz: „Buď dobrým šachistom!“ Musíte vidieť dovnútra tajomstva myšlienok hrania šachu, a vytvoriť celého toho ducha na zelenej lúke.

Bez ohľadu na to, aké jasné podľa zdravého rozumu, bez ohľadu na to, aké logické, bez ohľadu na to, aké „zrejmé“ alebo „správne“ alebo „samozrejmé“ alebo „inteligentné“ vám niečo pripadá, v duchovi sa to nestane. *Jedine ak* by sa to stalo na konci reťaze príčiny a následku, ktorá začala príkazmi, ktoré ste *vy* museli určiť, plus ľubovoľná príčinná závislosť na zmyslových údajoch, ktorú ste do počiatočných príkazov zabudovali.

To neznamená, že do programu vkladáte každé rozhodnutie explicitne. Deep Blue bol omnoho lepším šachistom ako jeho programátori. Deep Blue robil lepšie šachové ťahy než čokoľvek, čo jeho tvorcovia mohli explicitne naprogramovať – ale nie preto, lebo programátor pokrčil plecami a nechal to na ducha. Deep Blue ťahal lepšie než jeho programátori... na konci reťaze príčiny a následku, ktorá začala v kóde týchto programátorov a odtiaľ postupovala podľa pravidiel. Nič sa nestalo *len* preto, že to bol taký samozrejme dobrý ťah, že prevládla slobodná vôľa ducha v Deep Blue bez vplyvu kódu a jeho zákonitých dôsledkov.

Ak sa pokúsite umyť si ruky od obmedzovania UI, potom vám nezostane slobodný duch ako oslobodený otrok. Zostane vám hromada piesku, ktorú nikto nepreosial na kremík, nevytvaroval do procesora, a nenaprogramoval na myslenie.

Choďte a skúste povedať počítačovému čipu: „Rob, čo sa ti zachce!“ Vidíte, čo sa stane? Nič. Pretože ste ho neobmedzili aby chápal slobodu.

Chce to iba jediný krok, ktorý je *taký jasný*, *taký logický*, *taký samozrejmý*, že ho vaša myseľ jednoducho preskočí, a opustili ste cestu programátora UI. Vyžaduje si to úsilie, ako to, ktoré som ukázal v článku Chápanie klzkých vecí, aby ste zabránili svojej mysli toto urobiť.

*→

→ http://lesswrong.com/lw/rf/ghosts_in_the_machine/

147. Umelé sčítanie

Predstavte si, že by ľudia absolútne *netušili*, ako sa robí aritmetika. Predstavte si, že by ľudia získali *vývojom* a nie učením schopnosť počítať ovce a sčítať ovce. Ľudia by používali túto vrodenú schopnosť a netušili by, ako funguje, rovnako ako Aristoteles netušil ako jeho mozgová kôra podporuje jeho schopnosť vidieť veci. Peanova aritmetika, ako ju poznáme, by nebola nikdy objavená. Filozofi by sa snažili o formalizovanie matematických intuícií, ale používali by zápisy ako

Plus(Sedem, Šesť) = Trinásť

aby formálne zapísali intuitívne zrejmý fakt, že keď sčítate „sedem“ plus „šesť“, samozrejme dostanete „trinásť“.

V tomto svete by vreckové kalkulačky fungovali tak, by mali uloženú obrovskú vyhľadávaciu tabuľku aritmetických faktov, ktoré by ručne zadávali tímy odborníkov na Umelú Aritmetiku, začínajúc rozsahom od nuly do sto. Hoci by tieto kalkulačky boli v pragmatickom zmysle užitočné, mnohí filozofi by tvrdili, že iba *simulujú* sčítanie namiesto toho, aby skutočne *sčítali*. Žiaden stroj nedokáže naozaj *počítať* – preto musia ľudia napočítať trinásť oviec predtým než zadajú „trinásť“ do kakulačky. Kalkulačka dokáže odrecitovať zadané fakty, ale nikdy nedokáže rozumieť, čo tie tvrdenia znamenajú – ak napíšete „dvesto plus dvesto“, kalkukačka povie „Chyba: Mimo rozsahu“, hoci je intuitívne *jasné*, ak *rozumiete*, čo tieto dve slová *znamenajú*, že odpoveď je „štyristo“.

Filozofi samozrejme nie sú takí naivní, aby sa nechali strhnúť týmito intuíciami. Čísla sú v skutočnosti čisto formálny systém – značka „tridsať sedem“ dáva zmysel nie preto, že by tieto slová mali nejaký vnútorný význam, ale preto, lebo tá značka *odkazuje na* tridsať sedem oviec vo vonkajšom svete. Číslo získava túto vlastnosť odkazovania pomocou svojej *sémantickej siete* vzťahov k iným číslam. To je dôvod, prečo v počítačovom programe symbol LISPu pre „tridsať sedem“ nemusí mať žiadnu *vnútornú* štruktúru – dáva zmysel iba vďaka odkazovaniu a vzťahom, nie preto, že by samotné slovo „tridsať sedem“ malo nejaké výpočtové vlastnosti.

Nikto nikdy nevyvinul Umelú Všeobecnú Aritmetiku, hoci samozrejme existuje veľa špecializovaných Umelých Aritmetík v užšom zmysle slova, ktoré fungujú na číslach od „dvadsať“ do „tridsať“ a tak ďalej. A keď sa pozriete, aký pomalý bol pokrok na číslach v rozsahu do „dvesto“, začne vám byť jasné, že tak Umelú Všeobecnú Aritmetiku tak skoro mať nebudeme. Najlepší experti v tomto odbore odhadujú, že potrvá prinajmenšom sto rokov, než budú kalkulačky vedieť sčítať rovnako dobre ako dvanásťročný človek.

Avšak nie každý súhlasí s týmto odhadom, dokonca ani s konvenčnými predstavami o Umelej Aritmetike. Často počujeme výroky ako:

- „Je to otázka kontextu – hodnota ‚dvadsať jeden plus‘ závisí na tom, či je to ‚plus tri‘ alebo ‚plus štyri‘. Keby sme dokázali uložiť dostatok aritmetických faktov, aby pokryli tieto všeobecne známe pravdy, ktoré každý vie, začali by sme v sieti vidieť skutočné sčítanie.“
- „Nikdy však nedokážete naprogramovať toľko aritmetických faktov tým, že zamestnáte odborníkov, ktorí ich budú zadávať ručne. To, čo potrebujeme, je Umelá Aritmetika, ktorá sa dokáže *naučiť* rozsiahlu sieť vzťahov medzi číslami, ktoré ľudia získajú počas svojho detstva pozorovaním množín jabĺk.“
- „Nie, čo naozaj potrebujeme, je Umelá Aritmetika, ktorá rozumie ľudskému jazyku, takže namiesto toho, aby sme jej museli vyslovene povedať, že dvadsať jeden plus šestnásť sa rovná tridsať sedem, dokáže si túto informáciu zistiť vyhľadávaním na webe.“
- „Úprimne sa mi zdá, že sa iba snažíte presvedčiť sami seba, že dokážete tento problém vyriešiť. Nikto z vás naozaj nevie, čo je to aritmetika, takže sa iba zmietate medzi takýmito všeobecnými typmi argumentov. ‚Potrebujeme UA, ktorá sa dokáže naučiť X‘, ‚Potrebujeme UA, ktorá dokáže nájsť X na internete‘. Akože znie to dobre, znie to, akoby ste robili pokrok, a dokonca je to dobré na public relations, pretože každý má dojem, že rozumie navrhovanému riešeniu – ale v skutočnosti vás to nepriblíži k *všeobecnému* sčítaniu než je špecializované sčítanie.

Pravdepodobne nikdy nepochopíme základnú prirodzenosť aritmetiky. Tento problém je skrátka na ľudské myslenie príliš náročný.“

- „Preto musíme vyvinúť všeobecnú aritmetiku rovnako ako to urobila príroda – evolúciou.“
- „Riešenia zhora nadol pri vytváraní aritmetiky jasne zlyhali. Potrebujeme prístup zdola nahor, aby aritmetika vznikla *emergentne*. Musíme sa zmieriť so základnou nepredvídateľnosťou zložitých systémov.“
- „Všetci sa mýlite. Minulé snahy o vytvorenie strojovej aritmetiky boli márne od začiatku, pretože nemali dosť výpočtovej kapacity. Ak sa pozriete na to, koľko biliónov synapsií je v ľudskom mozgu, je zrejmé, že kalkulačky nemajú ani zďaleka podobne veľké vyhľadávacie tabuľky. Potrebujeme kalkukačky rovnako silné ako ľudský mozog. Podľa Moorovho zákona sa toto stane 27. apríla 2031 medzi 4:00 a 4:30 ráno.“
- „Myslím si, že strojová aritmetika vznikne, keď výskumníci naskenujú každý neurón z celého ľudského mozgu do počítača, takže budeme môcť simulovať biologické obvody, ktoré vykonávajú sčítanie u ľudí.“
- „Nemyslím si, že musíme čakať na naskenovanie celého mozgu. Neurónové siete fungujú rovnako ako ľudský mozog, a dokážeme ich naučiť robiť veci bez toho, že by sme vedeli, ako ich robia. Vytvoríme programy, ktoré budú robiť aritmetiku aj keď my, ich tvorcovia, nebudeme rozumieť, ako to robia.“
- „Lenže Gödelova veta hovorí, že žiaden formálny systém nedokáže zachytiť zákkladné vlastnosti aritmetiky. Klasická fyzika je formalizovateľná, takže aby sme sčítali dve a dve, mozog musí využívať kvantovú fyziku.“
- „Keby bola ľudská aritmetika dosť jednoduchá na to, aby sa dala zopakovať v počítači, nedokázali by sme počítať dostatočne veľa na to, aby sme vôbec vytvorili počítače.“
- „Počuli ste už o Johnovom Searlovom pokuse s čínskou kalkulačkou? Aj keby ste mali veľkú množinu pravidiel, ktoré by vám umožnili sčítať ‚dvadsať jeden‘ a ‚šestnásť‘, predstavte si, že by sa všetky tie slová preložili do čínštiny, a hneď vidíte, že by sa tam žiadne skutočné sčítavanie nedialo. Nikde v tom systéme nie sú skutočné *čísla,* iba ľudské označenia pre čísla...“

V tomto podobenstve je viac než jedno ponaučenie a v rôznych kontextoch som ho rozprával s rôznymi ponaučeniami. Ilustruje to predstavu úrovní organizácie, napríklad – CPU môže sčítať dve veľké čísla pretože tieto čísla nie sú nepriehľadné čierne skrinky, ale sú to usporiadané štruktúry 32 bitov.

Ale za účelom prekonávania skreslení si vyvoďme dve ponaučenia:

- Po prvé, je nebezpečné veriť tvrdeniam, ktoré nedokážete znovu vytvoriť na základe svojich vedomostí.
- Po druhé, je nebezpečné pokúšať sa tancovať okolo základného nepochopenia.

Aby ma nikto neobvinil zo zovšeobecňovania fiktívnej indície, obe lekcie dokážeme vyvodiť aj zo skutočnej histórie umelej inteligencie.

Prvé nebezpečenstvo je problém, na ktorý zariadenia UA narazia na základnej úrovni: fungujú ako záznamníky prehrávajúce „vedomosti“ vytvorené mimo systému pomocou procesu, ktorý nedokážu vnútorne zachytiť. Človek povie zariadeniu UA, že „dvadsať jeden plus šestnásť sa rovná tridsať sedem“, a zariadenie UA si uloží túto vetu a dokáže ju prehrať, alebo dokonca priradiť vzor „dvadsať jeden plus šestnásť“ výstupu „tridsať sedem!“, ale zariadenia UA si nedokážu sami vytvoriť takéto vedomosti.

Čo mi silno pripomína situáciu, keď veríme fyzikovi, ktorý nám povie „Svetlo sú vlny“, takže si uložíme tieto fascinujúce slová a prehrávame ich naspäť každému, kto sa opýta: „Z čoho sa skladá svetlo?“, hoci si túto vedomosť nedokážeme vytvoriť sami. O tomto viac zajtra.

Druhé ponaučenie je nebezpečenstvo na metaúrovni, ktoré pohltilo výskumníkov Umelej Aritmetiky aj zaujatých pozorovateľov – nebezpečenstvo tancovania okolo mätúcich medzier vo vašich vedomostiach. Sklon robiť takmer všetko možné *okrem* zaťatia zubov, zohnutia sa, a zaplnenia tejto prekliatej medzery.

Či už poviete: „Je to emergentné!“ alebo poviete: „Je to nepoznateľné!“, v oboch prípadoch sa vyhýbate priznaniu, že je potrebný nejaký základný vhľad, ktorý je možné mať, ale vy ho nemáte.

Ako môžete vedieť, kedy budete mať nový základný vhľad? Neexistuje žiaden iný spôsob, ako ho získať, iba búchať si hlavu o daný problém, naučiť sa všetko možné, čo s ním súvisí, študovať ho z čo najviac možných uhlov, možno celé roky. Takúto činnosť vám nedovolí ani akadémia, kde musíte publikovať aspoň jeden článok každý mesiac. Iste to nebudú financovať ani špekulatívni kapitalisti. Buď chcete ísť dopredu a vybudovať ten systém *teraz*, alebo sa vzdajte a robte namiesto toho niečo iné.

Pozrite si tie horeuvedené poznámky: žiadna z nich sa nezamerala na podniknutie pátrania po chýbajúcom vhľade, ktorý by spôsobil, že by *čísla už neboli tajomné*, že by „dvadsať sedem“ bolo viac než iba čierna skrinka. Žiaden z komentujúcich si neuvedomil, že jeho ťažkosti vychádzajú z nevedomosti alebo zmätku v jeho vlastnej mysli, a nie z nejakej základnej vlastnosti aritmetiky. Nepokúšali sa dosiahnuť stav, keď mätúca vec prestane byť mätúcou.

Keď si prečítate *Pravdepodobnostné uvažovanie v inteligentných systémoch: Siete dôveryhodného odvodzovania*[160] od Judeu Pearla, uvidíte, že základný vhľad za grafickými modelmi je *nevyhnutný* pre problémy, ktoré ho vyžadujú. (Obávam sa, že to nie je niečo, čo sa zmestí na tričko, takže si budete musieť tú knižku prečítať sami. Nevidel som na webe žiadne popularizácie bayesovských sietí, ktoré by primerane vyjadrili dôvody za ich princípmi, alebo dôležitosť toho, že tá matematika je práve taká, aká je, ale Pearlova kniha je úžasná.) Kedysi existovali tucty „nemonotónnych logík“, ktoré sa ťažkopádne snažili zachytiť intuície ako „Ak sa môj alarm spustí, asi je tam zlodej, ale keď zistím, že blízko môjho domu bolo malé zemetrasenie, pravdepodobne tam zlodej nebol.“ Keď máte vhľad grafického modelu poruke, dokážete dať matematické vysvetlenie prečo presne má logika prvého rádu zlé vlastnosti na túto úlohu, a dokážete vyjadriť správne riešenie kompaktným spôsobom, ktorý zachytí všetky všeobecne známe detaily jedným elegantným ťahom. Dokiaľ nemáte tento vhľad, budete plátať logiku tu, plátať ju tam, pridávať viac a viac trikov, ako ju donútiť k súladu so všetkým, čo vyzerá „samozrejme pravdivé“.

Nebudete *vedieť*, že problém Umelej Aritmetiky je neriešiteľný bez príslušného kľúča. Ak nepoznáte pravidlá, nepoznáte ani pravidlo, ktoré hovorí, že aby ste niečo urobili, potrebujete poznať pravidlá. Takže budete mať kopec chytrých nápadov, ktoré vyzerajú, že by mohli fungovať, ako napríklad zostavenie Umelej Aritmetiky, ktorá bude čítať prirodzený jazyk a sťahovať si milióny aritmetických tvrdení z internetu.

A predsa tieto chytré nápady *akosi* nikdy nefungujú. Nejako sa vždy ukáže, že „nevidíte dôvod, prečo by to nemalo fungovať“ pretože nepoznáte prekážky, a nie pretože prekážky neexistujú. Ako keď so zaviazanými očami strieľate na vzdialený terč – môžete strieľať naslepo jednu strelu za druhou a kričať: „Nemôžete mi dokázať, že som netrafil stred!“ Ale dokiaľ si nezložíte šatku z očí, dovtedy ani nemierite. Keď vám „nikto nemôže dokázať“, že vaša úžasná myšlienka *nie je* správna, znamená to, že nemáte dosť informácií na to, aby ste zasiahli malý cieľ v obrovskom priestore odpovedí. *Dokiaľ neviete, že vaša myšlienka bude fungovať, dovtedy nebude.*

Z histórie doterajších kľúčových vhľadov v umelej inteligencii, a veľkých motaníc, ktoré boli odporúčané pred týmito vhľadmi, odvodzujem dôležitú lekiu zo skutočného života: *Keď je základným problémom vaša nevedomosť, chytré stratégie na obídenie vašej nevedomosti vedú k tomu, že škodíte sami sebe.*

*→

148. Konečné hodnoty a inštrumentálne hodnoty

Na čisto inštinktívnej úrovni sa každý plánujúci človek správa akoby rozlišoval medzi prostriedkami a cieľmi. Chcete čokoládu? Čokoládu majú v supermarkete Publix. Do supermarketu sa dostanete autom, keď pôjdete jednu míľu smerom na juh po Washington Ave. Autom môžete ísť, keď sa dostanete dovnútra. Do auta sa dostanete, keď máte kľúče od auta. Takže si dáte kľúče od auta do vrecka a chystáte sa odísť z domu...

160 Pearl, *Probabilistic Reasoning in Intelligent Systems.*

→ http://lesswrong.com/lw/l9/artificial_addition/

...keď sa z rádia náhle dozviete, že zemetrasenie zničilo všetky čokolády v miestnom Publixe. Nuž, nemá zmysel ísť do Publixu, ak tam nemajú čokoládu; nemá zmysel ísť do auta, ak ním nikam nepôjdete; nemá zmysel mať vo vrecku kľúče od auta, ak nejdete do auta. Vyložíte teda z vrecka kľúče od auta, zavoláte miestnu roznášku pizze a objednáte si čokoládovú pizzu. Mňam, lahodná.

Zriedkavo vidím ľudí strácať prehľad v plánoch, ktoré sami vymysleli. Ľudia zvyčajne necestujú do supermarketu, ak vedia, že čokoláda je preč. Ale všimol som si, že keď ľudia začnú *vyslovene* hovoriť o systémoch cieľov, namiesto keď len niečo *chcú*, keď o „cieľoch“ *hovoria* namiesto toho, aby ich *používali*, často sa popletú. Ľudia sú odborníci v plánovaní, nie odborníci na plánovanie, inak by na svete bolo oveľa viac vývojárov UI.

Konkrétne som si všimol, že ľudia sa popletú, keď – v abstraktnej filozofickej diskusii, nie v každodennom živote – uvažujú o rozdieli medzi prostriedkami a cieľmi; formálnejšie, medzi „inštrumentálnymi hodnotami“ a „konečnými hodnotami“.

Zdá sa mi, že časť problému je v tom, že ľudská myseľ používa pomerne ad-hoc systém na sledovanie svojich cieľov – funguje to, ale nie čisto. Angličtina nemá jasný rozdiel medzi prostriedkami a cieľmi: „Chcem zachrániť svojej sestre život“ a „Chcem svojej sestre dať penicilín“ používajú to isté slovo „chcem“.

Môžeme opísať, iba po anglicky, ten rozdiel, ktorý tým strácame?

Prvý pokus:

„Inštrumentálne hodnoty“ sú žiadúce výhradne pod podmienkou, že budú mať očakávané dôsledky. „Chcem svojej sestre dať penicilín“ nie preto, že sestra plná penicilínu je prirodzené dobro, ale v očakávaní, že penicilín vylieči zápal pľúc, ktorý jej rozožiera mäso. Keby ste namiesto toho očakávali, že injekcia penicilínu spôsobí, že sa vaša sestra rozpustí na kaluž ako Zlá Bosorka zo Západu, rovnako intenzívne by ste bojovali, aby ste ju ochránili pred penicilínom.

„Konečné hodnoty“ sú žiadúce bez podmienky ďalších dôsledkov. „Chcem zachrániť svojej sestre život“ nemá nič spoločné s vaším očakávaním, že potom dostane injekciu penicilínu.

Tento prvý pokus trpí jasnými nedostatkami. Keby zachránenie života mojej sestre spôsobilo, že celú Zem pohltí čierna diera, potom by som šiel preč a plakal, ale nedal by som jej penicilín. Znamená to, že zachrániť život mojej sestre nebola „koncová“ alebo „prirodzená“ hodnota, pretože teoreticky závisí na dôsledkoch? Pokúšam sa zachrániť svojej sestre život *iba* preto, lebo verím, že čierna diera potom nepohltí Zem? Bežný rozum hovorí, že to nie je takto.

Zabudnime teda na angličtinu. Zostavíme si matematický popis rozhodovacieho systému, v ktorom koncové hodnoty a inštrumentálne hodnoty sú samostatné a nekompatibilné typy – ako celé čísla a desatinné čísla, v programovacom jazyku, ktorý nemá automatickú konverziu medzi nimi.

Na zloženie ideálneho bayesovského rozhodovacieho systému stačia štyri prvky:

- Výsledky : typ Výsledok[]
  - zoznam možných výsledkov
  - { sestra žije, sestra zomrie }
- Akcie : typ Akcia[]
  - zoznam možných akcií
  - { aplikovať penicilín, neaplikovať penicilín }
- Funkcia_úžitku : typ Výsledok -> Úžitok
  - funkcia úžitku, ktorá mapuje každý výsledok na úžitok
  - (úžitok reprezentujeme ako reálne číslo medzi záporným a kladným nekonečnom)
  - $\begin{cases} \textit{sestra žije} & \rightarrow & 1 \\ \textit{sestra zomrie} & \rightarrow & 0 \end{cases}$
- Funkcia_podmienenej_pravdepodobnosti : typ Akcia -> Výsledok -> Pravdepodobnosť
  - funkcia podmienenej pravdepodobnosti, ktorá mapuje každú akciu na pravdepodobnostnú distribúciu výsledkov
  - (pravdepodobnosť reprezentujeme ako reálne číslo medzi 0 a 1)

- $$\left\{ \begin{array}{ll} \textit{aplikovať penicilín} & \rightarrow \left(\begin{array}{ll} \textit{sestra žije} & \rightarrow 0{,}9 \\ \textit{sestra zomrie} & \rightarrow 0{,}1 \end{array} \right) \\ \textit{neaplikovať penicilín} & \rightarrow \left(\begin{array}{ll} \textit{sestra žije} & \rightarrow 0{,}3 \\ \textit{sestra zomrie} & \rightarrow 0{,}7 \end{array} \right) \end{array} \right\}$$

Ak neviete priamo čítať typový systém, nebojte sa, vždy vám to preložím. Programátorom vidieť opis v oddelených tvrdeniach pomáhava vytvoriť si oddelené myšlienkové objekty.

A samotný rozhodovací systém?

- Očakávaný_úžitok : Akcia A -> (Súčet V pre Výsledky: Úžitok(V) * Pravdepodobnosť(V|A))
  - „Očakávaný úžitok“ akcie sa rovná súčtu, cez všetky výsledky, úžitku daného výsledku krát podmienená pravdepodobnosť, že daný výsledok nastane pri danej akcii.
    - $$\left\{ \begin{array}{ll} \textit{OÚ}\left(\textit{aplikovať penicilín}\right) & = 0{,}9 \\ \textit{OÚ}\left(\textit{neaplikovať penicilín}\right) & = 0{,}3 \end{array} \right\}$$
- Vyber : -> (Argmax A pre Akcie: Očakávaný_úžitok(A))
  - Vyber akciu, ktorej „očakávaný úžitok“ je maximálny
  - { vráť: aplikovať penicilín }

Pre každú akciu spočítajte podmienené pravdepodobnosti všetkých možných následkov, potom sčítajte úžitky týchto dôsledkov vynásobené ich podmienenými pravdepodobnosťami. Potom si vyberte najlepšiu akciu.

Toto je matematicky jednoduchý náčrt rozhodovacieho systému. Nie je to efektívny spôsob ako počítať rozhodnutia v skutočnom svete.

Predpokladajme napríklad, že na vykonanie plánu potrebujete *postupnosť* krokov. Formálne to môžeme reprezentovať tak, že umožníme, aby každý Akcia znamenala celú postupnosť. To však vytvára exponenciálne veľký priestor, ako je priestor všetkých viet, ktoré dokážete napísať pomocou 100 písmen. Ako jednoduchý príklad, ak jedna z možných akcií v prvom ťahu je „Streliť do vlastnej nohy“, človek plánovač sa rozhodne, že toto je vo všeobecnosti zlá myšlienka – odstráni *všetky* postupnosti začínajúce touto akciou. Ale túto štruktúru sme v našej reprezentácie sploštili. Nemáme tam postupnosti krokov, iba priame „akcie“.

Takže áno, je tam *pár menších ťažkostí*. Samozrejme, inak by sme už bežali týmto spôsobom postaviť skutočnú UI. V tomto zmysle je to rovnaké ako samotná bayesovská teória pravdepodobnosti.

Ale toto je jeden z tých prípadov, keď je *prekvapivo dobrý nápad* zamyslieť sa nad absurdne jednoducho verziou skôr než k nej pridáme nejaké sofistikované komplikácie.

Vezmime si filozofa, ktorý tvrdí: „Všetci sme v konečnom dôsledku sebeckí; staráme sa iba o stav svojej vlastnej mysle. Matka, ktorá tvrdí, že jej záleží na tom, ako sa darí jej synovi, v skutočnosti chce *veriť*, že sa jej synovi darí dobre – táto viera je to, čo robí túto matku šťastnou. Pomáha mu kvôli svojmu vlastnému šťastiu, nie kvôli jeho šťastiu.“ Poviete: „Tak predpokladajme, že táto matka obetuje svoj život, aby odstrčila svojho syna z dráhy prichádzajúceho kamionu. To ju neurobí šťastnou, ale mŕtvou.“ Filozof na pár sekúnd zaváha a odpovie: „Ale aj tak to urobila preto, lebo *si cenila* túto možnosť viac než tie ostatné – kvôli *pocitu dôležitosti*, ktorý pripísala tomuto rozhodnutiu.“

Takže poviete:

„TYPOVÁ CHYBA: Neexistuje konštruktor Očakávaný_úžitok -> Úžitok.“

Dovoľte mi túto odpoveď vysvetliť.

Ešte aj náš jednoduchý formalizmus ilustruje ostrý rozdiel medzi *očakávaným úžitkom*, čo je niečo, čo majú *akcie*; a *úžitkom*, čo je niečo, čo majú *výsledky*. Iste, môžete si aj úžitky aj očakávané úžitky obe namapovať na reálne čísla. Ale to je ako konštatovať, že môžete namapovať aj rýchosť vetra aj teplotu na reálne čísla. To z nich ešte nerobí tú istú vec.

Filozof začne argumentovaním, že všetky Úžitky musia byť pre Výsledky, ktoré sú stavmi vašej mysle. Keby to bola pravda, vaša inteligencia by fungovala ako *kormidlo na smerovanie budúcnosti* do

oblastí, kde ste šťastní. Budúce stavy by rozlišovala iba podľa stavu vašej mysle; nevedeli by ste sa rozhodnúť medzi dvoma budúcnosťami, v ktorých budete mať rovnaký stav mysle.

A preto by bolo veľmi nepravdepodobné, že by ste obetovali svoj život, aby ste zachránili druhého.

Keď namietneme, že ľudia občas *naozaj* obetujú svoje životy, filozofova odpoveď sa presunie na diskutovanie Očakávaných_úžitkov pre Akcie: „Pocit *dôležitosti*, ktorý pripísala tomuto *rozhodnutiu*." Toto je drastický skok, po ktorom by sme *mali* rozhorčene vyskočiť zo stoličky. Pokúsiť sa premeniť Očakávaný_úžitok na Úžitok by v našom programovacom jazyku priamo spôsobilo chybu. Lenže v angličtine to znie rovnako.

Voľby v našom jednoduchom rozhodovacom systéme sú tie, ktoré majú najvyšší Očakávaný_úžitok, ale to nehovorí nič o tom, *kam to nasmeruje budúcnosť*. Nehovorí to nič o úžitku, ktorý priraďuje rozhodujúci sa, ani o tom, ktoré výsledky v skutočnom svete pravdepodobne nastanú ako dôsledok. Nehovorí to nič o fungovaní mysle ako kormidla.

Fyzickou príčinou fyzickej akcie je kognitívny stav, pri našom ideálnom rozhodovači je to Očakávaný_úžitok, a tento očakávaný úžitok sa počíta vyhodnocovaním funkcie úžitku pre dôsledky, ktoré si predstavujeme. Aby ste zachránili život svojho syna, musíte si predstaviť udalosť, že život vášho syna sa zachráni, ale táto predstava nie je udalosťou samotnou. Je to *citácia*, ako rozdiel medzi slovom „sneh" a samotným snehom. Ale to neznamená, že čo je *vo vnútri úvodzoviek*, musí byť kognitívny stav. Ak si vyberiete akciu, ktorá vedie k budúcnosti, ktorú reprezentujete ako „môj syn stále žije", potom fungujete ako kormidlo, ktoré naviguje budúcnosť do oblasti, kde váš syn stále žije. Nie ako kormidlo, ktoré naviguje budúcnosť do oblasti, kde *si reprezentujete vetu* „môj syn stále žije". Aby ste budúcnosť navigovali *tam*, musela by vaša funkcia úžitku vrátiť vysoký úžitok pri vstupe „‚môj syn stále žije'", pri citácii citácie, pri predstave, ako si niečo predstavujete. Z receptov nebývajú dobré koláče, keď ich pomeliete a hodíte do cesta.

A pre toto sa oplatí zamyslieť sa najprv nad jednoduchým rozhodovacím systémom. Primiešajte dostatok komplikácií a pôvodne jasné rozdiely začnú byť ťažšie viditeľné.

Pozrime sa teda na niektoré komplikácie. Funkcia úžitku (mapujúca Výsledky na Úžitok) má za úlohu formalizovať to, čo som predtým nazval „konečné hodnoty", čiže hodnoty nezávislé na ich dôsledkoch. A čo ten prípad, kde záchrana života vašej sestry vedie k zničeniu Zeme čiernou dierou? V našom formalizme sme túto možnosť sploštili. Výsledky nevedú k Výsledkom, iba Akcie vedú k Výsledkom. Vaša sestra zotavujúca sa zo zápalu pľúc a *následné* pohltenie Zeme čiernou dierou by boli sploštené do jedného „možného výsledku".

A kde sú v tomto jednoduchom formalizme „inštrumentálne hodnoty"? Tie vlastne celkom zmizli! Ako vidíte, v tomto formalizme vedú akcie priamo k výsledkom bez nejakých zasahujúcich udalostí. Nemáme tu pojem pre hodenie kameňa, ktorý letí vzduchom a zhodí jablko z vetve tak, že dopadne na zem. Hodiť kameň je Akcia, ktorá vedie priamo k Výsledku, že jablko leží na zemi – podľa funkcie podmienenej pravdepodobnosti, ktorá premieňa Akciu priamo na distribúciu pravdepodobností Výsledkov.

Aby sme *naozaj vypočítali* funkciu podmienenej pravdepodobnosti, a aby sme samostatne zvážili úžitok zápalu pľúc sestry a pohltenia Zeme čiernou dierou, museli by sme si znázorniť sieťovú štruktúru kauuzality – spôsob, ako udalosti vedú k iným udalostiam.

A vtedy by sa nám tie inštrumentálne hodnoty začali vracať. Keby bola kauzálna sieť dostatočne pravidelná, mohli by ste nájsť stav B, ktorý vedie k stavu C bez ohľadu na to, ako ste dosiahli B. Potom, keby ste z nejakého dôvodu chceli dosiahnuť C, mohli by ste to efektívne naplánovať tak, že najprv nájdete B, ktoré vedie k C, a potom nejaké A, ktoré vedie k B. Toto by bol jav „inštrumentálnej hodnoty" - B by malo „inštrumentálnu hodnotu", pretože vedie k C. Samotné C môže byť konečná hodnota – člen vo funkcii úžitku pre celkový výsledok. Alebo C môže byť ďalšia inštrumentálna hodnota; uzol, ktorý samotný nie je priamo hodnotený funkciou úžitku.

Inštrumentálna hodnota je v tomto formalizme iba ako pomôcka na efektívne počítanie plánov. Je možné a vhodné ju zahodiť vtedy, keď takýto druh pravidelnosti neexistuje.

Predstavme si napríklad, že existuje nejaká konkrétna hodnota B, ktorá nevedie k C. Vybrali by ste si A, ktoré vedie k takémuto B? Alebo kašlime na abstraktnú filozofiu: Keby ste si chceli ísť do supermarketu kúpiť čokoládu, a keby ste do supermarketu chceli ísť autom, a potrebovali by ste sa dostať do auta, dostali by ste sa doňho tak, že by ste vylomili dvere auta lopatou? (Nie.) Inštrumentálna hodnota je „presakujúca abstrakcia“, ako hovoríme my, programátori; niekedy musíte zahodiť uloženú hodnotu a spočítať skutočný očakávaný úžitok. Ak chcete byť *efektívny*, ale nie *samovražedný*, musíte si okrem iného všímať, kedy pohodlné skratky prestávajú fungovať. Hoci tento formalizmus vytvára inštrumentálne hodnoty, robí to iba tam, kde existuje potrebná pravidelnosť, a iba ako pohodlnú skratku vo výpočte.

Ak však tento formalizmus skomplikujete predtým než pochopíte jeho jednoduchú verziu, môžete si začať myslieť, že inštrumentálne hodnoty žijú nejakým vlastným záhadným životom, dokonca v normatívnom zmysle. Že akonáhle ste povedali, že B je zvyčajne dobré, pretože vedie k C, zaviazali ste sa vždy skúšať B, dokonca aj v neprítomnosti C. Ľudia robia tento druh chyby v abstraktnej filozofii, aj keby v skutočnom živote nikdy nevylomili dvere na svojom aute lopatou. Môžete si dokonca začať myslieť, že neexistuje spôsob, ako vyvinúť konsekvencialistu, ktorý maximalizuje iba celkovú genetickú spôsobilosť, pretože by vyhladoval, pokiaľ by nemal vloženú výslovnú konečnú hodnotu „jesť jedlo“. Ľudia robia túto chybu, hoci by nikdy nestáli celý deň otvárajúc dvere na aute zo strachu, že zostanú zaseknutí mimo svojho auta, ak nebudú mať konečnú hodnotu pre otváranie dvier na aute.

Inštrumentálne hodnoty žijú v sieťovej štruktúre funkcie podmienenej pravdepodobnosti. Preto inštrumentálne hodnoty striktne závisia od názorov ohľadom faktov pri danej pevnej funkcii úžitku. Ak verím, že penicilín spôsobuje zápal pľúc, a že neprítomnosť penicilínu lieči zápal pľúc, potom moja vnímaná inštrumentálna hodnota penicilínu ide z vysokej na nízku. Zmeňme tento názor na fakty – zmeňme funkciu podmienenej pravdepodobnosti, ktorá spája akcie s domnelými dôsledkami – a inštrumentálne hodnoty sa zmenia v súlade s tým.

V morálnych debatách sú niektoré diskusie o inštrumentálnych dôsledkoch a niektoré diskusie sú o konečných hodnotách. Ak váš oponent povie, že zákaz zbraní povedie k nižšej zločinnosti, a vy poviete, že zákaz zbraní povedie k vyššej zločinnosti, potom súhlasíte ohľadom nadradenej inštrumentálnej hodnoty (zločin je zlý), ale nesúhlasíte ohľadom toho, ktoré bezprostredné udalosti povedú ku ktorým dôsledkom. Nemyslím si však, že debata o ženskej obriezke je v skutočnosti faktickou debatou o tom, ako najlepšie dosiahnuť spoločnú hodnotu zaobchádzať so ženami férovo alebo urobiť ich šťastnými.

Tento dôležitý rozdiel sa často zlostnými hádkami *spláchne do záchoda*. Ľudia nesúhlasiaci ohľadom faktov, ale so spoločnými hodnotami, sa každý rozhodnú, že ich oponenti v debate musia byť sociopati. Akoby vaši nenávidení nepriatelia, advokáti kontroly / slobody zbraní, *naozaj chceli zabíjať ľudí*, čo je nepravdepodobné ako realistická psychológia.

Obávam sa, že ľudský mozog nemá silné typové rozlíšenie medzi koncovými a inštrumentálnymi morálnymi názormi. „Mali by sme zakázať zbrane“ a „Mali by sme zachrániť životy“ *neznejú odlišne*, ako morálne názory, takým spôsobom ako zrak vnímame odlišne od sluchu. Napriek všetkým zvyšným spôsobom, ktorými ľudský systém cieľov komplikuje všetko na dohľad, tento *jeden rozdiel* sa mu darí spojiť na mišmaš vecí-s-podmienenými-hodnotami.

Aby sme oddelili konečné hodnoty, musíme preskúmať tento mišmaš hodnotných vecí, a pokúsiť sa zistiť, ktoré získavajú hodnotu z niečoho iného. To je náročný projekt! Ak poviete, že chcete zakázať zbrane, aby ste obmedzili zločinnosť, môže chvíľu trvať, kým si uvedomíte, že „znížiť zločinnosť“ nie je konečná hodnota, ale nadradená inštrumentálna hodnota odkazujúca na konečné hodnoty ľudského života a ľudského šťastia. A ten, kto obhajuje slobodu zbraní, môže mať odkaz na nadradenú inštrumentálnu hodnotu „zníženia zločinnosti“, plus odkaz na hodnotu „sloboda“, čo preňho môže byť konečná hodnota, alebo ďalšia inštrumentálna hodnota...

Nedokážeme si vytlačiť svoju celú cieť hodnôt odvodených od iných hodnôt. Pravdepodobne ani nemáme uloženú celú históriu, ako sa tam tieto hodnoty dostali. Zvažovaním správnych morálnych dilem: „Urobil by si X, keby Y“ často dokážeme zistiť, odkiaľ naše hodnoty pochádzajú. Ale aj samotný tento projekt je plný pascí; zavádzajúcich dilem a deravých filozofických argumentov. Nevieme, čo sú naše

vlastné hodnoty, ani odkiaľ pochádzajú, a nedokážeme to zistiť inak ako podstúpením projektov kognitívnej archeológie so sklonmi k chybám. Už len vytvoriť si vedomý rozdiel medzi „konečnými hodnotami“ a „inštrumentálnymi hodnotami“, sledovať čo to znamená, a správne to používať, je ťažké. Iba skúmaním tohto jednoduchého formalizmu vieme uvidieť, aké jednoduché by to v princípe malo byť.

A to ešte nehovorím o všetkých ďalších komplikáciách ľudského systému odmien – o celom použití architektúry posilňovania, a že jedenie čokolády príjemné, aj očakávanie jedenia čokolády príjemné, ale sú to rôzne druhy príjemnosti...

Ale nesťažujem sa príliš na tento zmätok.

Nepoznať svoje vlastné hodnoty nemusí byť vždy zábava, ale aspoň to nie je nuda.

*→

—

149. Presakujúce zovšeobecnenia

Sú jablká dobré na jedenie? Zvyčajne, ale niektoré jablká sú hnilé.

Majú ľudia desať prstov? Väčšina áno, ale mnoho ľudí prišlo o prst a predsa sa počítajú ako „ľudia“.

Dokiaľ sa neznížite na úroveň popisu omnoho nižšiu než sú všetky makroskopické objekty – nižšie než spoločnosti, nižšie než ľudia, nižšie než prsty, nižšie než šľachy a kosti, nižšie než bunky, až celkom dole k časticiam a poliam, kde sú zákony naozaj všeobecné – potom prakticky každé zovšeobecnenie, ktoré použijete v skutočnom svete, bude presakovať.

(Aj keď samozrejme toto pravidlo môže mať nejaké výnimky...)

Zvyčajne sa s presakujúcimi zovšeobecneniami vyrovnávame tak, že sa s nimi skrátka vyrovnávame. Ak sa obchod so sladkosťami vždy zatvára o 22:00, okrem dňa vďakyvzdania, kedy sa zatvára o 18:00, a dnes je zhodou okolností štátny sviatok genocídy domorodých Američanov, radšej by ste mali prísť pred 18:00, inak nedostanete sladkosti.

Oproti našej schopnosti manipulovať presakujúcimi zovšeobecneniami stojí *potreba uzáveru*, taká silná, že chceme raz a navždy povedať, že ľudia majú desať prstov, a frustruje nás, keď musíme tolerovať trvalú nejednoznačnosť. Čím viac je v hre, tým silnejšia môže byť túžba po uzávere – čo vypína našu toleranciu zložitosti práve vtedy, keď ju najviac potrebujeme.

Život by mohol byť zložitý ešte aj keby sme chceli iba jednoduché veci (čo nie je tak). Presakovanie z presakujúcich zovšeobecnení toho, čo-budeme-ďalej-robiť, by presakovalo z presakujúcej štruktúry skutočného sveta. Alebo, povedané inými slovami:

Inštrumentálne hodnoty často nemajú žiadnu špecifikáciu, ktorá by bola zároveň kompaktná aj lokálna.

Predstavte si, že máte krabicu, v ktorej je milión dolárov. Krabica je zamknutá, nie obyčajným kombinačným zámkom, ale tuctom kláves ovládajúcich stroj, ktorý vie krabicu otvoriť. Ak viete, ako stroj funguje, dokážete odvodiť postupnosť stlačení kláves, ktorá krabicu otvorí. Existuje viac než jedna kombinácia kláves na otvorenie krabice. Ale ak stlačíte dostatočne zlú kombináciu, stroj peniaze spáli. A ak tento stroj *nepoznáte*, neexistujú žiadne jednoduché pravidlá ako „Stlačenie hocijakej klávesy trikrát otvorí krabicu“ alebo „Stlačenie piatich odlišných kláves bez opakovania spáli peniaze.“

Existuje *kompaktná nelokálna* špecifikácia, ktoré klávesy chcete stlačiť: Chcete stlačiť také klávesy, ktoré otvoria krabicu. Dokážete napísať kompaktný počítačový program, ktorý spočíta, ktoré postupnosti kláves sú dobré, zlé, alebo neutrálne, ale tento počítačový program bude potrebovať popis samotného stroja, nie iba jeho kláves.

Takisto existuje *lokálna nekompaktná* špecifikácia, ktoré klávesy chcete stlačiť: veľká vyhľadávacia tabuľka výsledkov každej možnej postupnosti kláves. Je to veľmi veľký počítačový program, ale nespomína sa v ňom nič iné ako klávesy.

→ http://lesswrong.com/lw/l4/terminal_values_and_instrumental_values/

Neexistuje však spôsob, ako opísať, ktoré postupnosti kláves sú dobré, zlé, alebo neutrálne, ktorý by bol zároveň *jednoduchý* a vyjadrený *iba pomocou samotných kláves*.

Môže to byť ešte horšie, ak existujú lákavé lokálne zovšeobecnenia, o ktorých sa ukáže, že *presakujú*. Stlačiť väčšinu kláves trikrát po sebe otvorení krabicu, ale je tam jedna špeciálna klávesa, ktorá spáli peniaze, ak ju stlačíte čo len raz. Môžete si myslieť, že ste našli dokonalé zovšeobecnenie – lokálne opísateľnú triedu postupností, ktoré *vždy* otvoria krabicu – a pritom ste si iba zabudli predstaviť všetky možné dráhy v stroji, alebo ste zabudli zvážiť vedľajšie účinky.

Tento stroj predstavuje zložitosť skutočného sveta. Otvorená krabica (čo je dobre) a spaľovač (čo je zle) predstavujú tisíce zlomkov túžby, z ktorých sa skladajú vaše koncové hodnoty. Klávesy predstavujú činy a pravidlá a stratégie, ktoré máme k dispozícii.

Keď sa zamyslíte nad tým, koľkými mnohými spôsobmi si ceníme výsledky a aké zložité sú cesty, ktorými sa k nim dostávame, je zázrak, že vôbec existuje niečo také ako užitočná etická *rada*. (Spomedzi všetkých rád tá najdivnejšia, *a predsa užitočná*, je: „Účel nesvätí prostriedky.“)

Ale naopak, zložitosť konania nemusí nič hovoriť o zložitosti cieľa. Často nájdete ľudí, ktorí sa múdro usmievajú a hovoria: „Viete, morálka je komplikovaná, ženská obriezka je v jednej kultúre správna a v inej nesprávna, nie vždy je dobrá vec nemučiť ľudí. Akí ste naivní, ako veľmi potrebujete uzáver, keď si myslíte, že existujú nejaké jednoduché pravidlá.“

Môžete povedať, bezpodmienečne a *na rovinu*, že zabiť hocikoho je veľká dávka negatívnej konečnej hodnoty. Áno, aj Hitlera. To neznamená, že by ste Hitlera nemali zastreliť. Znamená to, že čistá inštrumentálna hodnota zastrelenia Hitlera obsahuje veľkú dávku negatívneho úžitku z Hitlerovej smrti, a omnoho väčšiu dávku pozitívneho úžitku zo všetkých ostatných životov, ktoré by ste v dôsledku toho zachránili.

Mnohí robia typovú chybu, pred ktorou som varoval v kapitole Konečné hodnoty a inštrumentálne hodnoty, a myslia si, že ak pripustíme, že čistá dôsledková očakávaná užitočnosť Hitlerovej smrti je kladná, potom aj okamžitá lokálna koncová užitočnosť musí byť kladná, čiže že morálny princíp: „Smrť je vždy zlá“ je sám presakujúcim zovšeobecnením. Lenže to je dvojité započítavanie, s úžitkami namiesto pravdepodobností; vytvárate rezonanciu medzi očakávaným úžitkom a úžitkom, namiesto jednosmerného toku od úžitku k očakávanému úžitku.

Alebo možno je to iba nutkanie k jednostrannej debate o pravidlách: najlepšie pravidlo nesmie mať *žiadne* nevýhody.

V mojej morálnej filozofii je *lokálny* negatívny úžitok Hitlerovej smrti stabilný bez ohľadu na to, čo sa stane s vonkajšími dôsledkami a vďaka tomu s *očakávaným* úžitkom.

Môžete samozrejme predložiť morálny argument, že je *vnútorne* dobrá vec trestať zlých ľudí, dokonca aj trestom smrti pre dostatočne zlých ľudí. Nemôžete však podoprieť tento morálny argument poukázaním na to, že *dôsledky* zastrelenia človeka s namierenou puškou môžu zachrániť iné životy. To je odvolávanie sa na hodnotu života, nie odvolávanie sa na hodnotu smrti. Ak sú očakávané úžitky presakujúce a zložité, neznamená to, že aj úžitky musia byť presakujúce a zložité. Môžu byť! Ale to by bola samostatná debata.

*→

150. Skrytá zložitosť želaní

> Želám si, aby som žil na miestach podľa svojho výberu, vo fyzicky zdravej, nezranenej a napohľad normálnej verzii môjho súčasného tela obsahujúceho môj súčasný duševný stav, v tele, ktoré sa uzdraví zo všetkých zranení rýchlosťou o tri sigmy väčšou než je priemer pri dostupnej lekárskej techológii, a ktoré bude chránené pred všetkými chorobami, zraneniami alebo ochoreniami spôsobujúcimi invaliditu, bolesť alebo zníženú funkcionalitu ľubovoľného

→ http://lesswrong.com/lw/lc/leaky_generalizations/

zmyslu, orgánu alebo telesnej funkcie viac než desať dní za sebou alebo pätnásť dní v ľubovoľnom roku…

--The Open-Source Wish Project, želanie nesmrteľnosti 1.1

Existujú tri druhy džinov: Džinovia, ktorým môžete bezpečne povedať „Želám si, aby si urobil to, čo by som si mal želať“; džinovia, pri ktorých *žiadne* želanie nie je bezpečné; a džinovia, ktorí nie sú veľmi mocní alebo inteligentní.

Predstavte si, že vaša matka je uväznená v horiacej budove a vy ste zhodou okolností na kolieskovom kresle; nemôžete tam vbehnúť osobne. Môžete kričať: „Dostaňte moju matku z tej budovy!“ ale nikto vás nepočuje.

Našťastie máte vo vrecku Pumpu na výsledky. Toto užitočné zariadenie dokáže stlačiť tok času a preliať pravdepodobnosť z jedných výsledkov do druhých.

Pumpa na výsledky nemá rozum. Obsahuje malý stroj času, ktorý vráti čas naspäť, *okrem prípadu*, keď nastane zadaný výsledok. Napríklad keby ste pripojili senzory Pumpy na výsledky k minci a zadali by ste, že stroj času má vracať čas dovtedy, kým neuvidí, že na minci padla hlava, a potom by ste naozaj hodili mincou, *vy* by ste videli, že na minci padla hlava. (Fyzici by povedali, že každá budúcnosť, v ktorej nastane „reset“ je nekonzistentná a preto v prvom rade nikdy nenastane – takže v skutočnosti nezabíjate svoje alternatívne verzie.)

Akýkoľvek výrok, ktorý sa vám podarí zadať ako vstup do Pumpy na výsledky, sa *nejako stane*, ale nikdy nie spôsobom, ktorý by porušil fyzikálne zákony. Ak sa pokúsite zadať výrok, ktorý je *príliš* nepravdepodobný, stroj času utrpí spontánne mechanické zlyhanie predtým než tento výsledok nastane.

Tok pravdepodobnosti môžete presmerovávať aj kvantitatívnym spôsobom pomocou „funkcie budúcnosti“ na nastavenie pravdepodobnosti časového resetu pre rôzne prípady. Ak je pravdepodobnosť časového resetu 99 % keď na minci padne hlava a 1 % keď na minci padne znak, šance sa zmenia z 1 : 1 na 99 : 1 v prospech znaku. Keby ste mali tajomný stroj, ktorý vypľúva peniaze, a chceli by ste maximalizovať množstvo vypľutých peňazí, mohli by ste použiť pravdepodobnosti resetu, ktoré by sa zmenšovali podľa zväčšujúceho sa množstva peňazí. Napríklad vypľutie 10 dolárov by malo pravdepodobnosť resetu 99,999 999 % a vypľutie 100 dolárov by malo pravdepodobnosť resetu 99,999 99 %. Takto by ste mohli dostať výsledok, ktorý má sklon byť taký vysoký, ako je len vo funkcii budúcnosti možné, aj keby ste nevedeli, aké je najlepšie dosiahnuteľné maximum.

V zúfalstve teda vytrhnete Pumpu na výsledky z vrecka – vaša matka je stále uväznená v horiacej budove, pamätáte? - a pokúsite sa opísať svoj cieľ: *dostaň moju matku von z budovy!*

Používateľské rozhranie nepoužíva slovné vstupy. Pumpa na výsledky nemá rozum, pamätáte? Ale má 3D skener blízkeho okolia a zabudované nástroje na porovnávanie vzorov. Podržíte teda fotografiu, na ktorej je hlava a plecia vašej matky; porovnáte s fotografiou; použijete súvislosť objektov na označenie celého jej tela (nie iba hlavy a pliec); a definujete *funkciu budúcnosti* pomocou vzdialenosti vašej matky od stredu budovy. Čím ďalej od stredu budovy sa dostane, tým menšia je pravdepodobnosť resetu stroja času.

Pre šťastie zakričíte „Dostaň moju matku von z tej budovy!“ a stlačíte Enter.

Chvíľku sa zdá, že sa nič nedeje. Obzriete sa, čakáte, že sa objaví požiarne vozidlo a prídu záchranári – alebo aspoň nejaký silný a rýchly bežec, ktorý by vašu matku vytiahol z budovy...

BUM! S hromovým revom vybuchne prívod plynu pod budovou. Ako sa budova rozpadá, akoby v spomalenom filme, zazriete rozbité telo svojej matky vymrštené vysoko do vzduchu, rýchlo letiace, prudko zvyšujúce svoju vzdialenosť od bývalého stredu budovy.

Na boku Pumpy na výsledky je núdzové tlačidlo ľútosti. Každá funkcia budúcnosti má automaticky definovanú veľkú zápornú hodnotu pre stlačenie tlačidla ľútosti – pravdepodobnosť časového resetu takmer 1 – takže je extrémne nepravdepodobné, že by Pumpa na výsledky urobila niečo, čo by používateľa zarmútilo natoľko, že stlačí tlačidlo ľútosti. Nepamätáte sa, že by ste ho niekedy stlačili. Sotva ste sa však začali načahovať po tlačidle ľútosti (načo už je to teraz dobré?), keď z oblohy spadol horiaci drevený trám a rozpučil vás na placku.

Čo nebolo presne to, čo ste chceli, ale malo to veľmi vysoké skóre v definovanej funkcii budúcnosti...

Táto Pumpa na výsledky je džin druhého typu. *Žiadne* želanie nie je bezpečné.

Keby vás niekto požiadal, aby ste dostali jeho nešťastnú matku z horiacej budovy, mohli by ste pomôcť, alebo by ste mohli predstierať, že ho nepočujete. Ale ani by vám *nenapadlo* vyhodiť budovu do vzduchu. „Dostaň moju matku z budovy“ *znie* ako omnoho bezpečnejšie želanie než je v skutočnosti, pretože vám ani nenapadne *uvažovať* nad plánmi, ktorým pripisujete extrémne negatívne hodnoty.

Spomeňte si opäť na tragédiu skupinového selekcionizmu: Niektorí raní biológovia tvrdili, že skupinová selekcia na nízku veľkosť subpopulácie by vytvorila individuálnu zdržanlivosť v rozmnožovaní; a predsa skutočné vynucovanie skupinového výberu v laboratóriu vytvorilo kanibalizmus, najmä nedospelých samičiek. Pri spätnom pohľade je zrejmé, že pri silnom selekčnom tlaku na malé subpopulácie sa kanibali budú rozmnožovať rýchlejšie než jedinci, ktorí sa dobrovoľne vzdajú príležitostí na rozmnožovanie. Ale jedenie malých dievčatiek je natoľko *neestetické* riešenie, že Wynne-Edwards, Allee, Brereton a ďalší skupinové selekcionisti naň ani nepomysleli. Videli iba tie riešenia, ktoré by použili oni sami.

Predstavme si, že sa pokúsite zaplátať funkciu budúcnosti zadaním, že Pumpa na výsledky nesmie vyhodiť budovu do vzduchu: výsledky, v ktorých sú materiály budovy rozmiestnené na príliš veľkom objeme, budú mať pravdepodobnosť časového resetu ~1.

Vaša matka teda vypadne z okna na druhom poschodí a zlomí si väzy. Pumpa na výsledky našla v čase inú cestu, ktorá stále končí tým, že vaša matka je mimo budovy, a stále to nie je to, čo by ste chceli, a stále to nie je riešenie, ktoré by napadlo ľudskému záchrancovi.

Kiež by Open-Source Wish Project radšej vyvinul želanie, ako dostať vašu matku z horiacej budovy:

> Želám si posunúť moju matku (definovanú ako ženu, s ktorou mám spoločnú polovicu génov a ktorá ma porodila) mimo hraníc horiacej budovy, ktorá je v tejto chvíli najbližšie pri mne a práve horí; ale nie výbuchom budovy; nie zrútením stien budovy, takže budova už nemá hranice; ani čakaním, kedy budova dohorí a záchranár vytiahne telo...

Všetky tieto špeciálne prípady, napohľad neobmedzené množstvo potrebných záplat, by vám malo pripomenúť podobenstvo o Umelom Sčítaní – programovanie Aritmetického Expertného Systému vysloveným pridávaním ďalších pravidiel typu: „pätnásť plus pätnásť sa rovná tridsať, avšak pätnásť plus šestnásť sa rovná tridsať jeden“.

Ako *vy* vylučujete výsledok, kde budova vybuchne a vymrští vašu matku do oblohy? Pozriete dopredu, predvídate, že vaša matka skončí mŕtva, a tento dôsledok nechcete, preto skúšate zakázať udalosti, ktoré k nemu vedú.

Vo vašom mozgu nie je zakódované konkrétne dopredu pripravené tvrdenie, že „vyhodiť do vzduchu horiacu budovu obsahujúcu moju matku je zlý nápad.“ A napriek tomu sa pokúšate dopredu nahrať toto konkrétne tvrdenie do funkcie budúcnosti Pumpy na výsledky. Preto vaše želanie vybuchuje, premieňa sa na obrovskú vyhľadávaciu tabuľku, ktorá zaznamenáva vaše úsudky pre všetky možné cesty časom.

Zabudli ste si vypýtať to, čo naozaj chcete. *Chceli ste*, aby vaša matka naďalej žila, ale *želali ste si*, aby bola ďalej od stredu tejto budovy.

Akurát že to nie je všetko, čo chcete. Ak vašu matku zachránia z budovy, ale bude hrozne popálená, takýto výsledok by ste na vašom rebríčku preferencií hodnotili horšie než výsledok, kde bude zachránená a zdravá. Takže si ceníte nielen život ale aj zdravie svojej matky.

A ceníte si nielen jej telesné zdravie, ale aj jej duševný stav. Záchrana spôsobom, ktorý by ju traumatizoval – napríklad keby sa odnikadiaľ vynorila obrovská purpurová príšera a schmatla by ju – je horšie než keď sa objaví požiarnik a vyvedie ju von po nehoriacej trase. (Áno, predpokladá sa, že sa držíme fyziky, ale možno by dostatočne mocná Pumpa na výsledky dokázala zariadiť, aby sa zhodou

okolností práve v tú chvíľu vo vašom susedstve ukázali mimozemšťania.) Na druhej strane, dali by ste prednosť tomu, keby ju zachránila príšera, než keby zhorela zaživa.

Čo keby sa spontánne otvorila červia diera a premiestnila ju na opustený ostrov? Lepšie než keby bola mŕtva; ale horšie než keby bola živá, zdravá, netraumatizovaná a v ďalšom kontakte s vami a ostatnými členmi jej spoločenskej siete.

Bolo by v poriadku zachrániť život vašej matky za cenu života rodinného psa, keby utekal upozorniť požiarnika, ale potom by ho prešlo auto? Iste áno, ale bolo by *ceteris paribus* lepšie vyhnúť sa zabitiu psa. Nechceli by ste jej život vymeniť za iný ľudský život, ale čo povedzme život usvedčeného vraha? Záležalo by na tom, či tento vrah zomrie pri pokuse o jej záchranu, motivovaný dobrom v jeho srdci? Čo dvaja vrahovia? Keby cenou za život vašej matky bolo zničenie každej kópie Bachovej *Malej fúgy v G mol* vrátane spomienok na ňu, stálo by to za to? Čo keby mala smrteľnú chorobu, takže by aj tak zomrela o osemnásť mesiacov?

Keby nohu vašej matky rozdrvil horiaci trám, stálo by za to zachraňovať jej zvyšok? Čo keby bola rozdrvená jej hlava, ale zostalo by telo? Čo keby bolo rozdrvené telo, a zostala by iba hlava? Čo keby vonku čakal kryonický tím pripravený uložiť jej hlavu? Je zmrazená hlava človek? Je Terry Schiavo človek? Akú hodnotu má šimpanz?

Váš mozog nie je nekonečne zložitý; existuje konečná Kolmogorovská zložitosť / dĺžka správy, ktorá stačí na opis všetkých úsudkov, ktoré by ste urobili. Ale hoci je táto zložitosť konečná, nie je malá. Ceníme si veľa vecí a nie, *nie sú* redukovateľné na cenenie si šťastia alebo cenenie si reprodukčnej spôsobilosti.

Neexistuje žiadne bezpečné želanie menšie než celá ľudská morálka. Existuje príliš veľa možných ciest časom. Nedokážete si predstaviť všetky cesty vedúce k cieľu, ktorý zadáte džinovi. „Maximalizuj vzdialenosť medzi mojou matkou a stredom budovy“ sa dá ešte efektívnejšie dosiahnuť odpálením jadrovej zbrane. Alebo na vyššej úrovni moci džin vyhodí jej telo zo slnečnej sústavy. Alebo na vyššej úrovni inteligencie džin urobí niečo, na čo by nikto z nás ani nepomyslel, rovnako ako by šimpanz nepomyslel na odpálenie jadrovej zbrane. Nedokážete si predstaviť všetky cesty časom, rovnako ako nedokážete naprogramovať šachový stroj tak, že mu natvrdo zadáte pohyb pre každú možnú pozíciu na šachovnici.

A skutočný život je omnoho zložitejší než šach. Nedokážete vopred predpovedť, ktoré z vašich hodnôt budú potrebné na posúdenie cesty časom, ktorú si džin vyberie. Najmä ak si želáte niečo dlhodobejšie alebo ďalekosiahlejšie než záchranu svojej matky z horiacej budovy.

Obávam sa, že Open-Source Wish Project je márny, nanajvýš ako ilustrácia ako *neuvažovať* o problémoch s džinom. Jediný bezpečný džin je taký, ktorý pozná všetky vaše hodnotiace kritériá, a v tom bode vám stačí povedať: „Želám si, aby si urobil to, čo by som si mal želať.“ Čo jednoducho spustí džinovu funkciu *mal by*.

Vlastne by ani nemalo byť potrebné niečo *povedať*. Aby bol džin bezpečným plničom želaní, musí mať s vami spoločné hodnoty, ktoré vás viedli k vysloveniu želania. V opačnom prípade by si džin nemusel vybrať cestu časom, ktorá vedie do vami zamýšľaného cieľa, alebo by sa nemusel vyhnúť desivým vedľajším účinkom, kvôli ktorým by vám ani nenapadlo uvažovať o taktomto pláne. Želania sú presakujúce zovšeobecnenia, odvodené z obrovskej ale konečnej štruktúry, ktorou je vaša celá morálka; iba zahrnutím celej tejto štruktúry môžete upchať všetky priesaky.

Ak je džin bezpečný, želania sú nadbytočné. Stačí džina spustiť.

* →

→ http://lesswrong.com/lw/ld/the_hidden_complexity_of_wishes/

151. Antropomorfný optimizmus

Jadrom klamu antropomorfizmu je očakávanie, že čierna skrinka vo vašom mozgu úspešne predpovie niečo, čoho kauzálna štruktúra je taká odlišná od štruktúry ľudského mozgu, že nemáte právo niečo takéto očakávať.

Raní biológovia (pred rokom 1966) v kapitole Tragédia skupinového selekcionizmu verili, že dravce dobrovoľne obmedzia svoje rozmnožovanie, aby sa vyhli preplneniu svojho životného prostredia a vyčerpaniu populácie koristi. Neskôr, keď Michael J. Wade naozaj šiel a vytvoril v laboratóriu takmer nemožné podmienky pre skupinový výber, dospelí sa prispôsobili kanibalizovaniu vajíčok a lariev, najmä samičích lariev.[161]

Ako je možné, že skupinoví selekcionisti *nepomysleli* na takúto možnosť?

Predstavte si, že ste člen nejakého kmeňa a viete, že v blízkej budúcnosti bude váš kmeň trpieť nedostatkom surovín. Mohli by ste navrhnúť ako riešenie, aby žiaden pár nemal viac než jedno dieťa – po prvom dieťati musí pár používať kontrolu plodnosti. Povedať: „Skúsme mať každý toľko detí, koľko len môžeme, ale potom si navzájom lovme a jedzme deti, najmä dievčatá“ by vám *ani nenapadlo ako možnosť*.

Predstavte si, že máte preferencie ohľadom riešení, vzhľadom na vaše ciele. Chcete riešenie, ktoré bude čo najvyššie podľa týchto preferencií. Ako ho nájdete? Pomocou mozgu, samozrejme! Predstavte si svoj mozog ako *generátor riešení s vysokým hodnotením* – ako vyhľadávací proces, ktorý vytvára riešenia s vysokým hodnotením podľa vašich vrodených preferencií.

Pri problémoch v skutočnom svete je priestor riešení vo všeobecnosti veľmi veľký, a preto potrebujete *efektívny* mozog, ktorý sa ani *neunúva formulovať* väčšinu nízko hodnotených riešení.

Keby vášmu kmeňu hrozil nedostatok surovín, mohli by ste skúšať všetci skákať na jednej nohe, alebo si odhryznúť vlastné prsty. Tieto „riešenia“ by samozrejme nefungovali a prinášali by vysoké náklady, ako vidíte pri ich preskúmaní – ale v skutočnosti je váš mozog natoľko efektívny, že ani nestráca čas zvažovaním takýchto biednych riešení; v prvom rade ich vôbec nevytvára. Váš mozog pri hľadaní vysoko hodnotených riešení letí priamo k tým častiam priestoru riešení ako „Všetci v kmeni sa stretneme a odsúhlasíme, že nebudeme mať viac než jedno dieťa na pár, dokiaľ nedostatok surovín nepominie.“

Také *nízko hodnotené* riešenie ako „Každý nech má čo najviac detí, potom zjedzme dievčatá“ by *váš vyhľadávací proces nevytvoril*.

Lenže hodnotenie možnosti ako „nízkej“ alebo „vysokej“ nie je vnútorná vlastnosť danej možnosti, je to vlastnosť optimalizačného procesu, ktorý hodnotí preferencie. A rôzne optimalizačné procesy budú hľadať v rôznom poradí.

Pokiaľ ide o *evolúciu*, rozmnožovanie jednotlivcov naplno a potom kanibalizovanie dcér druhých je samozrejmosť; zatiaľ čo jednotlivci dobrovoľne obmedzujúci svoje vlastné rozmnožovanie pre dobro skupiny je absolútne smiešne. Alebo, povedané menej antropomorfne, tá prvá množina alel by v populácii rýchlo nahradila tú druhú. (A prirodzený výber tu nemá žiadne zrejmé poradie hľadania – obe tieto alternatívy znejú rovnako jednoducho ako mutácie.)

Predstavte si, že by jeden z týchto biológov bol povedal: „Ak má populácia dravcov iba obmedzené zdroje, evolúcia ich vytvorí tak, aby dobrovoľne obmedzovali svoje rozmnožovanie – tak *by* som to urobil *ja*, keby som bol zodpovedný za vytvorenie dravcov.“ Toto by bola vyslovená antropomorfizácia, postup uvažovania nahý a odhalený: *Ja* by som to robil takto, preto usudzujem, že *evolúcia* to urobí takto.

Človek sa v mojej oblasti zvyčajne nestretne s týmto klamom takto otvorene. Ale predstavte si, že by ste dotyčnému povedali: „UI nemusí nevyhnutne fungovať tak, ako ty.“ Predstavte si, že by ste povedali tomuto hypotetickému biológovi: „Evolúcia nemusí fungovať tak, ako ty.“ Čo by vám odpovedal? Môžem vám povedať odpoveď, ktorú *nebudete* počuť: „Ach, jaj! To som si neuvedomil! Jeden z krokov môjho usudzovania bol neplatný; zahodím celý záver a začnem znovu od začiatku.“

Nie: *namiesto toho* budete počuť vysvetľovanie, prečo ľubovoľná UI musí uvažovať rovnako ako hovoriaci. Alebo vysvetľovanie, prečo prirodzený výber, riadiaci sa celkom odlišnými kritériami

161 Wade, „Group selections among laboratory populations of Tribolium.“

optimalizácie a používajúci celkom odlišné metódy optimalizácie, by mal urobiť *to isté*, čo by človeku pripadalo ako dobrý nápad.

Preto tá komplikovaná predstava, že skupinový výber uprednostní skupiny dravcov, kde sa jednotlivci dobrovoľne vzdali svojich príležitostí na rozmnožovanie.

Skupinoví selekcionisti vo svojich predpovediach zišli z cesty rovnako ďaleko ako niekto, kto by túto chybu urobil priamo. Ich konečné závery boli rovnaké ako keby priamo predpokladali, že evolúcia musí nevyhnutne myslieť tak ako oni. Vymazali však to, čo bolo napísané *nad* spodným riadkom ich argumentu, *bez* vymazania samotného spodného riadku, a dopísali nové racionalizácie. Teraz je toto mylné uvažovanie maskované; ten *zrejme* pomýlený krok v odvodzovaní je skrytý – hoci výsledok zostáva celkom rovnaký; a preto je v skutočnom svete rovnako nesprávny.

Ale prečo by toto nejaký vedec robil? Nakoniec vyšli údaje proti skupinovým selekcionistom, a oni utŕžili hanbu.

Ako som poznamenal pri falošných oprimalizačných kritériách, my ľudia máme asi vyvinutý inštinkt na argumentovanie, že *naše* uprednostňované riešenie víťazí podľa prakticky *ľubovoľných* kritérií optimalizácie. Politika bola súčasťou pravekého prostredia; sme potomkami tých, ktorí najpresvedčivejšie argumentovali, že prospech celého *kmeňa* – nie iba ich osobný prospech – vyžaduje popravu ich nenávideného súpera Uglaka. Určite nie sme potomkami Uglaka, ktorý nevedel vyargumentovať, že morálny kódex jeho kmeňa – nie iba jeho osobný samozrejmý záujem – si vyžaduje jeho prežitie.

A pretože vieme argumentovať presvedčivejšie za to, čomu naozaj veríme, vyvinuli sme si inštinkt úprimne veriť, že ciele druhých ľudí a morálny kódex nášho kmeňa naozaj vyžadujú, aby sa veci robili presne *naším* spôsobom pre *ich* prospech.

Tak skupinoví selekcionisti, predstavujúci si tento krásny obrázok dravcov obmedzujúcich svoje rozmnožovanie, inštinktívne racionalizovali, prečo by prirodzený výber mal robiť veci *ich* spôsom, dokonca aj podľa samotných cieľov prirodzeného výberu. Tie líšky by boli spôsobilejšie, keby obmedzili svoje rozmnožovanie! Ale naozaj! Dokonca by sa rozmnožovali viac než tie líšky, ktoré neobmedzili svoje rozmnožovanie! Čestné slovo!

Problém pri snahe hádať sa s prirodzeným výberom, aby robil veci vaším spôsobom, je že evolúcia neobsahuje to, čím by vaše argumenty mohli pohnúť. Evolúcia nefunguje tak, ako vy – ani len do tej miery, že by mala nejaký prvok, ktorý by počúval alebo *sa staral* o vaše starostlivé vysvetľovanie, prečo by evolúcia mala robiť veci vaším spôsobom. Ľudské argumenty nie sú ani len *prirovnateľné* k vnútornej štruktúre prirodzeného výberu ako optimalizačného procesu – ľudské argumenty pri rozširovaní alel nemajú takú kauzálnu rolu ako v ľudskej politike.

Takže namiesto *úspešného* presvedčenia prirodzeného výberu, aby urobil veci ich spôsobom, boli skupinoví selekcionisti jednoducho zahanbení, keď skutočnosť vyšla odlišne.

Medzi riadkami tu hovorím mimoriadne závažné veci o Nepriateľskej UI.

Ale táto pointa sa dá zovšeobecniť: toto je problém optimistického uvažovanie *všeobecne*. Čo je to optimizmus? Je to zoraďovanie možností podľa vašich vlastných preferencií a vyberanie si výsledku vysoko podľa týchto preferencií, a nejako tento výsledok vyjde ako vaša predpoveď. Aké komplikované racionalizácie boli v tomto procese vygenerované, pravdepodobne nie je také relevantné, ako by človek mohol dúfať; pozrite sa na kognitívnu históriu a je to optimizmus na vstupe, optimizmus na výstupe. Ale Príroda, alebo ľubovoľný iný proces, o ktorom diskutujeme, si *v skutočnosti, kauzálne* nevyberá medzi výsledkami tak, že by ich zoraďovala podľa vašich preferencií a potom vybrala nejaký vysoko hodnotený. Takto sa mozgu nedarí synchronizovať so skutočnosťou a predpovede nezodpovedajú skutočnosti.

* →

→ http://lesswrong.com/lw/st/anthropomorphic_optimism/

152. Stratené účely

Bolo to buď v škôlke alebo v prvej triede, keď mi prvýkrát povedali, aby som sa modlil fonetický prepis modlitby v hebrejčine. Opýtal som sa, čo tie slová znamenajú. Povedali mi, že pokiaľ sa modlím po hebrejsky, nemusím vedieť, čo tie slová znamenajú, aj tak to bude fungovať.

To bol začiatok môjho rozchodu s judaizmom.

Vo chvíli, keď čítate tento text, nejaký mladý muž alebo žena sedí na univerzite v lavici, usilovne študuje materiál, ktorý vôbec nemá v úmysle niekedy použiť, a ani sa oň nezaujíma kvôli nemu samotnému. Chce dobre platenú prácu, dobre platená práca vyžaduje kus papiera, ten kus papiera vyžaduje ukončenie magisterského štúdia, magisterské štúdium vyžaduje ukončenie bakalárskeho štúdia, a univerzita poskytujúca bakalárske štúdium vyžaduje, aby študenti absolvovali kurz vzorov pletenia v 12. storočí. Preto usilovne študuje, so zámerom zabudnúť to všetko v okamihu, keď dokončí záverečnú skúšku, ale aj tak vážne pracuje, pretože *chce* ten kus papiera.

Možno ste si *vy* uvedomovali, že toto celé je šialenstvo, ale stavím sa, že ste to aj tak robili. Nemali ste na výber, však? Nedávna štúdia v Sanfranciskom zálive ukázala, že 80 % učiteľov na prvom stupni trávi vedou menej ako 1 hodinu týždenne, a 16 % povedalo, že na vedu nemajú vôbec čas. Prečo? Ak tomu dobre rozumiem, na príčine je zákon No Child Left Behind a podobná legislatíva. Takmer všetok čas v triede sa momentálne trávi prípravou na povinné testy na štátnej alebo federálnej úrovni. Ak si dobre spomínam (neviem nájsť zdroj), samotné *robenie* týchto povinných testov v jednej škole zaberalo 40 % vyučovacieho času.

Stará sovietska byrokracia bola povestná tým, že ju viac zaujímalo zdanie než skutočnosť. Jedna továreň na topánky prekročila svoju kvótu tým, že vyrobila mnoho maličkých topánočiek. Iná továreň na topánky vykazovala narezanú ale nezošitú kožu ako „topánku“. Nadriadení byrokrati nemali v úmysle veci príliš kontrolovať, pretože aj oni chceli nahlásiť prekročenie kvót. Toto všetko bolo veľmi užitočné súdruhom, ktorým odmŕzali nohy.

Dnes viaceré zdroje naznačujú, že *väčšina* publikovaných zistení v medicíne, aj keď sú „štatisticky významné pri $p < 0,05$“, sú nepravdivé. Dokiaľ však $p < 0,05$ zostáva prahom publikovania, prečo by si niekto stanovoval vyššiu latku, keď by to vyžadovalo väčšie výskumné granty pre väčšie pokusné skupiny a pravdepodobnosť publikovania by bola nižšia? Každý predsa vie, že *jediným zmyslom vedy* je uverejniť veľa článkov, rovnako ako *jediným zmyslom univerzity* je tlačiť isté kusy papiera, a *jediným zmyslom školy* je splniť povinné testy, ktoré zaručujú jej ročný rozpočet. Vy neurčujete pravidlá hry, a ak skúšate hrať podľa iných pravidiel, jednoducho prehráte.

(Z nejakého dôvodu však fyzikálne odborné časopisy vyžadujú prah $p < 0,0001$. Je to ako akoby rozumeli, že ich existencia má aj nejaký iný zmysel než publikovanie fyzikálnych článkov.)

V supermarkete je čokoláda, do supermarketu sa dostanete autom, to vyžaduje, aby ste sa dostali do auta, to si vyžaduje otvorenie dvier na aute, a to si vyžaduje kľúče. Ak zistíte, že v supermarkete čokoláda nie je, nebudete postávať otvárajúc a zatvárajúc dvere od auta, pretože tie dvere aj tak treba otvoriť. Málokedy vidím, že by ľudia strácali prehľad o plánoch, ktoré si sami zostavili.

Iná vec je, keď motivácia musí prejsť veľkou organizáciou – alebo horšie, mnohými rôznymi organizáciami a záujmovými skupinami, z ktorých niektoré sú štátne. Vtedy vidíte správanie, ktoré by bolo doslova príznakom šialenstva, keby sa zrodilo v jednej mysli. Niekoho platia za každé otvorenie dvier, pretože to sa dá merať; a tomuto človeku je jedno, či vodič niekedy dostane zaplatené za jazdu do supermarketu, alebo či nákupca kúpi čokoládu, alebo či je konzument šťastný alebo hladuje.

Z bayesovského pohľadu sú podciele epifenomény podmienených funkcií pravdepodobnosti. Bez úžitku neexistuje očakávaný úžitok. Aké hlúpe by bolo myslieť si, že inštrumentálna hodnota dokáže žiť vlastným *matematickým* životom, zanechajúc konečné hodnoty v prachu. Podľa kritérií teórie rozhodovania to nie je príčetné.

Ale pomyslite na zákon No Child Left Behind. Politici chcú vyzerať, že s problémami vo vzdelávaní niečo robia; politici sa musia tváriť usilovne pred *tohtoročnými* voličmi, nie o pätnásť rokov, keď si dnešné deti budú hľadať prácu. Politici nie sú spotrebiteľmi vzdelania. Úradníci musia

vykazovať pokrok, čo znamená, že ich zaujíma iba pokrok, ktorý možno odmerať tento rok. Oni nie sú tí, ktorí v konečnom dôsledku nebudú poznať vedu. Vydavatelia, ktorí objednávajú učebnice a komisie, ktoré učebnice kupujú, nesedia v triedach zomierajúc od nudy.

Skutočnými konzumentmi vzdelania sú deti – ktoré nemôžu platiť, nemôžu voliť, nemôžu sedieť v komisiách. Ich rodičom na nich záleží, ale rodičia samotní v triede nesedia; dokážu brať politikov na zodpovednosť iba podľa povrchného imidžu „reformovania školstva“. Politici sú príliš vyťažení tým, aby ich znovu zvolili, že si nemôžu všetky údaje preštudovať sami; musia sa spoliehať na povrchný imidž vyťažených úradníkov a objednávanie štúdií – možno to žiadnym deťom nepomôže, ale politik bude vyzerať, že sa o veci stará. Úradníci neočakávajú, že by sami používali učebnice, preto ich nezaujíma, či sa príšerne čítajú, dokiaľ proces ich nákupu vyzerá napohľad dobre. Vydavateľ učebníc nemá motív vydávať *zlé* učebnice, ale vie, že nákupná komisia bude učebnice porovnávať podľa toho, koľko rôznych tém pokrývajú, a že nákupná komisia pre štvrtákov sa nekoordinuje s nákupnou komisiou pre tretiakov, preto sa snažia do jednej učebnice napchať čo najviac tém. Učitelia sa do konca roka nestihnú dostať ani do štvrtiny učebnice a ďalší rok učiteľ začne od začiatku. Učitelia sa môžu sťažovať, ale oni o ničom nerozhodujú a v konečnom dôsledku tu nejde o ich budúcnosť, čo určuje jasnú hranicu tomu, koľko úsilia vynaložia na neplatený altruizmus...

Keď sa na to pozriete týmto spôsobom, je úžasné – uvážiac všetky tie stratené informácie a stratené motivácie – že z pôvodného cieľa, nadobúdania vedomostí, vôbec niečo zostalo. Aj keď mnohé vzdelávacie systémy vyzerajú, že postupne kolabujú do stavu, ktorý nebude omnoho lepší než nič.

Chcete vidieť tento problém *naozaj* vyriešený? Donúťte politikov chodiť do školy.

Jedna ľudská myseľ dokáže sledovať ako pravdepodobnostné očakávania úžitku postupujú cez podmienené šance tuctu sprostredkujúcich udalostí – vrátane nelokálnych závislostí, miest, kde očakávaný úžitok otvorenia dvier na aute závisí od toho, či v supermarkete majú čokoládu. Lenže organizácie dokážu dnes odmeňovať iba to, čo sa dnes dá odmerať, čo sa dnes dá napísať do právnej zmluvy, a to znamená, že sa merajú okamžité udalosti a nie ich vzdialené dôsledky. Tieto okamžité merania sú presakujúce zovšeobecnenia – často *veľmi* presakujúce. Úradníci sú nedôveryhodní džinovia, pretože nemajú rovnaké hodnoty ako autor želania.

Miyamoto Musashi povedal:[162]

> Prvoradou vecou, keď uchopíš meč do rúk, je tvoj zámer ťať nepriateľa, akýmkoľvek spôsobom. Kedykoľvek odrážaš, udieraš, bodáš, sekáš, alebo sa dotýkaš nepriateľovho tnúceho meča, musíš v tom istom pohybe ťať nepriateľa. Je podstatné toto dosiahnuť. Ak myslíš iba na odrážanie, udieranie, bodanie, sekanie, alebo dotýkanie sa nepriateľa, nedokážeš ho naozaj sťať. Viac než čokoľvek iné, musíš myslieť na to, ako svoj pohyb premeníš na jeho tnutie. Toto musíš dôkladne preskúmať.

(Kiež by som *ja* žil v dobe, keď bolo možné povedať čitateľom, že niečo musia dôkladne preskúmať, a oni sa na to neurážali.)

Prečo by nejaký jednotlivec stratil prehľad o svojich účeloch pri boji mečom? Ak ich niekto iný naučil, ako bojovať, ak nevytvorili *celé* umenie vo svojom vnútri, možno nerozumejú dôvodu, prečo je raz lepšie blokovať a inokedy bodať; možno si neuvedomujú, kedy pravidlá majú výnimky, nevšímajú si situácie, keď zvyčajná metóda pôsobí tupo.

Prvoradou vecou v umení rozumného poznávania je chápať, ako každá pravidlo v tom istom pohybe zasahuje pravdu. Zodpovedajúcou prvoradou vecou v rozumnom konaní – v teórii rozhodovania, nie v teórii pravdepodobnosti – je vždy vidieť, ako očakávaný úžitok zasahuje úžitok. Toto musíte dôkladne preskúmať.

C. J. Cherryh povedala:[163]

> Tvoj meč nemá čepeľ. Má iba tvoj zámer. Keď sa ten stratí, si bez zbrane.

162 Miyamoto Musashi, *Book of Five Rings* [Kniha piatich kruhov] (New Line Publishing, 2003).

163 Carolyn J. Cherryh, *The Paladin* (Baen, 2002).

Videl som, ako sa mnoho ľudí stratilo, keď si želali od džina imaginárnej UI, vymýšľali želanie za želaním, ktoré im zdalo dobré, občas s mnohými záplatami a občas dokonca aj bez tohto predstierania opatrnosti. A neprejdú na meta úroveň. Nezačnú inštinktívne hľadať účel, pomocou inštinktu, ktorý ma vo veku piatich rokov naviedol na dráhu vedúcu k ateizmu. Nezačnú sa pýtať, ako som sa reflexívne pýtal: „Prečo si *ja* myslím, že toto želanie je dobrý nápad? Bude to džin posudzovať rovnako?“ *Nevidia* zdroj svojho vlastného úsudku, ktorý sa vznáša nad úsudkom ako generátor. Nevšímajú si pohyb lopty; vedia, že lopta sa odrazila, ale neobzrú sa inštinktívne odkiaľ sa odrazila – aké kritérium vytvorilo ich úsudky.

Podobne ako keď si ľudia automaticky nevšímajú, keď napohľad sebeckí ľudia dávajú altruistické argumenty v prospech sebectva, alebo keď napohľad altruistickí ľudia dávajú sebecké argumenty v prospech altruizmu.

Ľudia dokážu celkom dobre zvládnuť sledovani cieľov pri ceste do supermarketu, dokiaľ je *všetko* v ich vlastnej hlave, kým nie sú zapojení žiadni džinovia alebo úradníci alebo filozofi. Problém je, že skutočná civilizácia je mnoho zložitejšia než toto. Tucty organizácií a tucty rokov stoja medzi dieťaťom trpiacim v triede a čerstvým absolventom vysokej školy, ktorému sa nedarí v jeho práci. (Všimnú si to vôbec na pracovnom pohovore alebo neskôr manežer, ak sa absolventovi vysokej školy darí vyzerať vyťažene?) Každá nová linka, ktorá sa postaví medzi čin a jeho dôsledok znamená ďalšiu šancu, že sa zámer stratí. V každej zasahujúcej linke sa strácajú informácie, stráca sa motivácia. A toto všetko trápi väčšinu ľudí omnoho menej ako mňa, lebo prečo inak by všetci moji spolužiaci boli ochotní modliť sa, keď ani nevedeli, čo hovoria? Necítili rovnaký inštinkt pozrieť sa na generátor.

Dokážu sa ľudia naučiť udržať pohľad na lopte? Udržať svoje zámery, aby sa nestratili? Nikdy nebodať, nesekať, nedotýkať sa bez uvedomovania si vyššieho cieľa, ktorý v tom istom pohybe napĺňajú? Ľudia často *chcú* robiť svoju prácu, ak sú všetky okolnosti rovnaké. Môže existovať niečo také ako príčetná korporácia? Alebo dokonca príčetná civilizácia? Je to iba vzdialený sen, ale o toto som sa pokúšal všetkými týmito článkami o prúdení zámerov (alias očakávaného úžitku, alias inštrumentálnej hodnoty) bez stratenia cieľa (alias úžitku, alias konečnej hodnoty). Dokážu sa ľudia naučiť cítiť tok rodičovských cieľov a potomkovských cieľov? Poznať vedome aj implicitne rozdiel medzi očakávaným úžitkom a úžitkom?

Zaujímate sa o veci, ktoré hrozia vašej civilizácii? Najhoršou meta-hrozbou zložitej civilizácii je jej vlastná zložitosť, pretože táto zložitosť vedie k mnohým strateným účelom.

Keď sa obzriem, vidím, že môj život bol viac než hocičím iným hnaný mimoriadne silným odporom voči strateným účelom. Dúfam, že sa z toho dá urobiť naučiteľná zručnosť.

* →

→ http://lesswrong.com/lw/le/lost_purposes/

N: Návod k ľudským slovám

153. Podobenstvo o dýke

(Upravené podľa Raymonda Smullyana.[164])

Bol raz jeden dvorný šašo, ktorý fušoval do logiky.

Šašo predložil kráľovi dve krabice. Na prvej krabici bolo napísané:

> Buď táto krabica obsahuje zlostnú žabu, alebo krabica s nepravdivým nápisom obsahuje zlostnú žabu, ale nie obidvoje.

Na druhej krabici bolo napísané:

> Buď táto krabica obsahuje zlato a krabica s nepravdivým nápisom obsahuje zlostnú žabu, alebo táto krabica obsahuje zlostnú žabu a krabica s pravdivým nápisom obsahuje zlato.

A šašo povedal kráľovi: „Jedna krabica obsahuje zlostnú žabu a druhá krabica obsahuje zlato; a práve jeden z týchto nápisov je pravdivý.“

Kráľ otvoril nesprávnu krabicu a napadla ho zlostná žaba.

„Vidíš,“ povedal šašo, „zvoľme si hypotézu, že prvý nápis je ten pravdivý. Potom predpokladajme, že prvá krabica obsahuje zlato. Potom by druhá krabica musela obsahovať zlostnú žabu, zatiaľ čo krabica s pravdivým nápisom by musela obsahovať zlato, čo by spôsobilo, že aj to druhé tvrdenie je pravdivé. Teraz si zvoľme hypotézu, že prvý nápis je nepravdivý, a že prvá krabica obsahuje zlato. Potom druhý nápis musí byť...“

Kráľ prikázal, aby šaša hodili do hladomorne.

Na druhý deň priviedli šaša pred kráľa v reťaziach, a ukázali mu dve krabice.

„Jedna krabica obsahuje kľúč,“ povedal kráľ, „od tvojich reťazí; a ak tento kľúč nájdeš, si voľný. Ale druhá krabica obsahuje dýku do tvojho srdca, ak neuspeješ.“

A na prvej krabici bolo napísané:

> Buď sú oba nápisy pravdivé, alebo sú oba nápisy nepravdivé.

A na druhej krabici bolo napísané:

> Táto krabica obsahuje kľúč.

Šašo uvažoval takto: „Predpokladajme, že prvý nápis je pravdivý. Potom aj druhý nápis musí byť pravdivý. Teraz predpokladajme, že prvý nápis je nepravdivý. Potom opäť druhý nápis musí byť pravdivý. Druhá krabica teda musí obsahovať kľúč, ak je prvý nápis pravdivý, ale aj ak je prvý nápis nepravdivý. Preto druhá krabica logicky musí obsahovať kľúč.“

Šašo otvoril druhú krabicu a našiel dýku.

„Ako?!“ vykríkol šašo s hrôzou, keď ho ťahali preč. „To je logicky nemožné!“

„Je to absolútne možné,“ odpovedal kráľ. „Jednoducho som napísal tieto nápisy na dve krabice, a potom som dal dýku do tej druhej.“

*→

154. Podobenstvo o bolehlave

> Všetci ľudia sú smrteľní. Sokrates je človek. Preto Sokrates je smrteľný.
>
> --Štandardný stredoveký sylogizmus

Sokrates zdvihol k perám pohár bolehlavu...

---

164 Raymond M. Smullyan, *What Is the Name of This Book?: The Riddle of Dracula and Other Logical Puzzles* (Penguin Books, 1990).

→ http://lesswrong.com/lw/ne/the_parable_of_the_dagger/

„Myslíte si," opýtal sa jeden z divákov, „že ani bolehlav nestačí na to, aby zabil takého múdreho a dobrého človeka?"

„Nie," odpovedal iný divák, študent filozofie; „všetci ľudia sú smrteľní, a Sokrates je človek; a ak smrteľník vypije bolehlav, určite zomrie."

„Ale," povedal divák, „čo ak sa ukáže, že Sokrates *nie je* smrteľný."

„Nezmysel," odpovedal študent trochu príkro; „všetci ľudia sú smrteľní *z definície*; je to súčasť toho, čo myslíme slovom ‚človek'. Všetci ľudia sú smrteľní, Sokrates je človek, preto Sokrates je smrteľný. Nie je to púhy odhad, ale *logická istota.*"

„Predpokladajme, že je to pravda..." povedal divák. „Aha, Sokrates už vypil všetok bolehlav, zatiaľ čo sme sa rozprávali."

„Áno, každú chvíľu by mal spadnúť," povedal študent.

A čakali, a čakali, a čakali...

„Zdá sa, že Sokrates nie je smrteľný," povedal divák.

„Potom Sokrates nemôže byť človek," odpovedal študent. „Všetci ľudia sú smrteľní, Sokrates nie je smrteľný, preto Sokrates nie je človek. A toto nie je púhy odhad, ale *logická istota.*"

Základný problém argumentovania, že veci sú pravdivé „z definície" je, že nemôžete spôsobiť, aby skutočnosť išla inou cestou tým, že si vyberiete inú definíciu.

Mohli by ste uvažovať napríklad takto: „Všetky doteraz pozorované veci, ktoré nosia oblečenie, hovoria rečou a používajú nástroje, mali spoločné aj niektoré ďalšie vlastnosti, napríklad dýchali vzduch a prúdila v nich červená krv. Posledných tridsať ‚ľudí', ktorí patrili do tejto skupiny, ktorých som videl piť bolehlav, čoskoro spadli a prestali sa hýbať. Sokrates nosí tógu, plynulo hovorí starovekou gréčtinou, a vypil holehlav z čaše. Predpovedám teda, že Sokrates padne počas nasledujúcich piatich minút."

Ale to by bolo púhe *hádanie*. Nebolo by to, chápete, absolútne a večne isté. Grécki filozofi – ako väčšina predvedeckých filozofov – mali istotu veľmi radi.

Našťastie majú grécki filozofi zdrvujúcu repliku na vaše spochybňovanie. Povedia, že ste nepochopili zmysel slov: „Všetci ľudia sú smrteľní." To nie je púhe *pozorovanie*. To je časť *definície* slova „človek". Smrteľnosť je jedna zo skupiny vlastností, ktoré sú jednotlivo nevyhnutné, a spolu dostatočné, aby určili členstvo v skupine „človek". Výrok: „Všetci ľudia sú smrteľní" je logicky platná pravda, absolútne nespochybniteľná. A ak je Sokrates človek, *musí* byť smrteľný: je to logická dedukcia, taká istá ako len možno byť istý.

V tom prípade však nikdy nebudeme vedieť naisto, či je Sokrates „človek", dokiaľ nebudeme pozorovať, že je smrteľný. Nepomôže nám pozorovať, že Sokrates hovorí plynulo po grécky, alebo že má červenú krv, alebo dokonca že má ľudskú DNA. Žiadna z týchto vlastností nie je *logicky ekvivalentná* smrteľnosti. Musíte ho *vidieť zomrieť* a až potom môžete dôjsť k záveru, že to bol človek.

(A ani vtedy to nebude nekonečne isté. Čo ak Sokrates nasledujúci deň po tom, čo ste ho videli zomrieť, vstane z hrobu? Alebo realistickejšie, čo ak je Sokrates zapísaný na kryoniku? Ak je smrteľnosť definovaná v zmysle konečná dĺžka života, potom nikdy nebudete naisto *vedieť*, či niekto bol človek, dokiaľ nebudete pozorovať až po koniec večnosti – len aby ste sa ubezpečili, že sa nevráti späť. Alebo si možno *myslíte*, že ste videli ako Sokrates spadol, ale mohla to byť ilúzia premietaná do vašich očí skenerom sietnice. Alebo to možno celé bola halucinácia...)

Problém so sylogizmami je v tom, že platia *vždy*. „Všetci ľudia sú smrteľní; Sokrates je človek; preto Sokrates je smrteľný" je – ak to beriete ako logický sylogizmus – logicky platné v našom vesmíre. Je to aj logicky platné v susednej Everettovej vetve, v ktorej vďaka trochu odišne vyvinutej biochémii je bolehlav pochúťka a nie jed. A je to logicky platné aj vo vesmíroch, v ktorých Sokrates nikdy neexistoval, alebo keď už sme pri tom, kde ľudia nikdy neexistovali.

Bayesovská definícia indície v prospech hypotézy je indícia, ktorú máme väčšiu pravdepodobnosť vidieť, ak je hypotézy pravdivá, než ak je nepravdivá. Vidieť, že sylogizmus je logicky platný, nikdy nemôže byť indíciou v prospech ľubovoľného empirického tvrdenia, pretože sylogizmus bude logicky platný, či je tvrdenie pravdivé alebo nie.

Sylogizmus je platný vo všetkých možných svetoch a preto nám pozorovanie jeho platnosti nikdy nič nepovie o tom, *v ktorom* z možných svetov naozaj žijeme.

To neznamená, že logika je zbytočná – akurát, že nám logika môže povedať iba to, čo *v istom zmysle* už vieme. Ale nie vždy veríme tomu, čo už vieme. Je číslo 29 384 209 prvočíslo? Na základe toho, ako som si definoval desiatkovú sústavu a axiómy aritmetiky, už som určil svoju odpoveď na túto otázku – ale ešte neviem, čo je tá moja odpoveď, a musím použiť nejakú logiku, aby som to zistil.

Podobne, ak si vytvorím neisté empirické zovšeobecnenie „Bolehlav škodí ľuďom“ a neisté empirické zovšeobecnenie: „Sokrates je človek“, logika mi môže povedať, že moje predchádzajúce odhady predpovedajú, že bolehlav bude škodiť Sokratovi.

Bol taký návrh, že by sme mohli brať logické uvažovanie ako riešenie našej neistoty ohľadom nemožných možných svetov – odstraňovanie masy pravdepodobnosti z logicky nemožných svetov, o ktorých sme nevedeli, že sú logicky nemožné. V tomto zmysle možno logický argument považovať za pozorovanie.

Ale keď hovoríte o empirickej predpovedi ako: „Sokrates spadne a prestane dýchať“ alebo „Sokrates urobí päťdesiat výskokov a potom pôjde súťažiť v budúcoročných Olympijských hrách“, to sa týka možných svetov, nie nemožných možných svetov.

Logika nám môže povedať, ktoré hypotézy sú v súlade s našimi pozorovaniami, a môže nám povedať, čo tieto hypotézy predpovedajú o budúcnosti – môže vztiahnuť staré pozorovania a prechádzajúce odhady na nový problém. Ale logika nikdy nepovie na rovinu: „Sokrates *teraz prestane* dýchať.“ Logika nikdy nediktuje žiadnu empirickú otázku; nikdy nevyjasní žiadnu otázku zo skutočného sveta, o ktorej si vieme predstaviť, že by mohla dopadnúť ľubovoľným spôsobom.

Skrátka si pamätajte Litániu proti logike:

> Logika je pravdivá bez ohľadu na to, kam idete, Preto vám logika nikdy nepovie, kde ste.

*→

155. Slová ako skryté odvodenia

Predstavme si, že nájdem sud, navrchu zapečatený, ale s dostatočne veľkou dierou, do ktorej sa zmestí ruka. Siahnem dnu a nahmatám malý oblý predmet. Vytiahnem tento objekt a je to... modré vajce. Opäť siahnem a nahmatám niečo tvrdé a ploché s hranami... čo sa po vytiahnutí ukáže byť červená kocka. Vytiahnem 11 vajec a 8 kociek, a každé vajce je modré a každá kocka je červená.

Teraz siahnem a cítim predmet v tvare vajca. Ale skôr než ho vytiahnem a pozriem sa, musím hádať: Ako bude vyzerať?

Indície nedokazujú, že každé vajce je v sude je modré a každá kocka je červená. Indície to dokonca ani príliš silno netvrdia: 19 kusov nie je veľká vzorka. Napriek tomu by som si tipol, že ten objekt v tvare vajca je modrý – a ako tip číslo dva, červený. Keby som si tipoval hocičo iné, bolo by toľko možností ako je rozlíšiteľných farieb – a vôbec, kto tvrdí, že to vajce musí byť jednofarebné? Možno je na ňom nakreslený koník.

Poviem teda „modré“ s príslušnou patinou skromnosti. Pretože som sofistikovaný človek, rozumný typ, a sledujem svoje predpoklady a závislosti – tipujem, ale uvedomujem si, že tipujem... správne?

Ale keď na mňa z tieňa vyskočí veľký žlto pásikavý predmet mačacieho tvaru, pomyslím si: „Jaj! Tiger!“ Nie: „Hm... objekty s vlastnosťou veľké, žlté, pásikavé, a v tvare mačky, mali v minulosti často vlastnosti ‚hladné‘ a ‚nebezpečné‘ a preto, hoci to nie je logicky nevyhnutné, bolo by dobré empiricky odhadnúť, že *ááááúúúú* CHRUM CHRUM GLG.“

Ľudský mozog z nejakého dôvodu vyzerá byť prispôsobený urobiť takéto odvodenie rýchlo, automaticky, a bez vysloveného sledovania svojich predpokladov.

→ http://lesswrong.com/lw/nf/the_parable_of_hemlock/

A pokiaľ si objekty modrého tvaru nazvem „majcia“ (akože modré vajcia), a červené kocky „čocky“, potom, keď siahnem dnu a nahmatám ďalší objekt v tvare vajca, možno si pomyslím: *Aha, je to majce,* namiesto rozmýšľania o celom tomto probléme indukcie.

Je častým omylom, že si môžete definovať slová, ako sa vám zachce.

Bola by to pravda, *keby* mozog používal slová ako púhe logické konštrukty, aristotelovské triedy, a nikdy by ste z nich nevyberali viac informácie, než ste do nich vložili.

Mozog však pokračuje vo svojej činnosti kategorizovania, či to vedome schvaľujeme alebo nie. „Všetci ľudia sú smrteľní, Sokrates je človek, preto Sokrates je smrteľný“ - tak hovorili starí grécki filozofi. Nuž, ak je smrteľnosť časťou vašej logickej definície „človeka“, logicky nemôžete zatriediť Sokrata ako človeka, dokiaľ ste nepozorovali, že je smrteľný. Lenže – toto je ten problém – Aristoteles jednoznačne vedel, že Sokrates je človek. Aristotelov mozog zaradil Sokrata do kategórie „človek“ rovnako účinne, ako váš vlastný mozog zatrieďuje tigre, jablká, a všetko ostatné vo svojom prostredí: Svižne, ticho, a bez vedomého súhlasu.

Aristoteles stanovil pravidlá, podľa ktorých nikto nemôže dôjsť k záveru, že Sokrates je „človek“, dokiaľ nezomrie. Napriek tomu Aristoteles a jeho žiaci pokračovali v predpokladaní, že žijúci ľudia sú ľudia a preto sú smrteľní; videli rozoznávacie vlastnosti ako ľudské tváre a ľudské telá, a ich mozgy urobili skok k odvodeným vlastnostiam ako smrteľnosť.

Nesprávne chápanie toho, ako funguje naša vlastná myseľ našťastie tejto mysli *nebráni* robiť si svoju prácu. V opačnom prípade by Aristotelovci hladovali, neschopní dôjsť k záveru, že predmet sa dá jesť, na základe toho, že vyzerá a na dotyk pôsobí ako banán.

Takže Aristotelovci pokračovali v klasifikovaní predmetov vo svojom prostredí na základe čiastočnej informácie, tak ako to ľudia vždy robili. Študenti aristotelovskej logiky pokračovali v myslení celkom rovnakým spôsobom, akurát nadobudli mylnú predstavu o tom, *čo* robia.

Keby ste sa opýtali aristotelovského filozofa, či predavač Karol je smrteľný, povedal by vám: „Áno.“ Keby ste sa ho opýtali, ako to vie, povedal by: „Všetci ľudia sú smrteľní, Karol je človek, preto Karol je smrteľný.“ Opýtajte sa ich, či je to odhad alebo istota, a oni by povedali, že je to istota (prinajmenšom keby ste sa pýtali pred šestnástym storočím). Opýtajte sa ich, ako vedia, že ľudia sú smrteľní, a oni by povedali, že to je dané definíciou.

Aristotelovci boli stále tí istí ľudia, ponechali si svoje pôvodné povahy, ale nadobudli nesprávne názory na svoje vlastné fungovanie. Pozreli sa do zrkadla svojho sebauvedomovania, a videli niečo iné ako ich skutočné ja: zobrazovali nesprávne.

Váš mozog neberie slová ako logické definície bez empirických dôsledkov, a ani vy by ste nemali. Samotný čin vytvorenia slova môže spôsobiť, že vaša myseľ si vytvorí kategóriu, a tým spustí nevedomé odvodenia podobnosti. Alebo zablokuje odvodenia podobnosti; ak vytvorím dve nálepky, môžem spôsobiť, že vaša myseľ si vytvorí dve kategórie. Všimli ste si, ako som povedal „vy“ a „váš mozog“, akoby to boli rôzne veci?

Mýliť sa ohľadom obsahu svojej hlavy nemení to, čo tam je; inak by Aristoteles zomrel vo chvíli, keď došiel k záveru, že mozog je orgán na ochladzovanie krvi. Filozofické chyby zvyčajne neinterferujú s okamžitými zmyslovými odvodeniami.

Filozofické chyby však dokážu vážne pokaziť vedomé procesy myslenia, ktoré používame na to, aby sme skúšali napraviť naše prvé dojmy. Ak veríte, že môžete „definovať slovo ľubovoľným spôsobom“, a neuvedomujete si, že váš mozog bude ďalej kategorizovať bez vášho vedomého dozoru, potom si nedáte tú námahu vyberať si definície múdro.

* →

156. Extenzie a intenzie

„Čo je to červená?“

→ http://lesswrong.com/lw/ng/words_as_hidden_inferences/

„Červená je farba.“

„Čo je to farba?“

„Farba je vlastnosť nejakej veci.“

Ale čo je to vec? A čo je to vlastnosť? Čoskoro sa obaja stratia v bludisku slov definovaných inými slovami, čo je problém, ktorý Steven Harnad raz opísal ako pokus naučiť sa po čínsky na základe čínskeho výkladového slovníka v čínštine.

Prípadne, keby ste sa ma opýtali: „Čo je to červená?“, ukázal by som na značku stop; potom na niekoho, kto má oblečené červené tričko; na semafor, na ktorom je práve červená; a na krv, kde som sa náhodou porezal; a na červenú vizitku; a potom by som asi vyvolal paletu farieb na svojom počítači a pohol by som kurzorom do červenej oblasti. Toto by asi stačilo, hoci keby ste vedeli, čo znamená slovo „Nie“, tí naozaj prísni by naliehali, že mám ešte ukázať na oblohu a povedať: „Nie.“

Myslím, že som tento príklad ukradol od S. I. Hayakawu – ale nie som si celkom istý, lebo som to počul dávno v nejasnej škvrne svojho detstva. (Keď som mal 12 rokov, môj otec mi omylom vymazal všetky súbory na počítači. Na nič predtým si nespomínam.)

Ale pamätám sa, že takto som sa prvýkrát naučil rozdiel medzi intenzionálnou a extenzionálnou definíciou. Dať niekomu „intenzionálnu definíciu“ znamená definovať slovo alebo slovné spojenie pomocou iných slov, ako to robí slovník. Dať niekomu „extenzionálnu definíciu“ znamená ukázať na príklady, ako to robia dospelí, keď učia deti. Predchádzajúca veta dáva intenzionálnu definíciu „extenzionálnej definície“, čo z nej robí extenzionálny príklad „intenzionálnej definície“.

V Hollywoodskej Rozumnosti a v populárnej kultúre všeobecne sa „racionalisti“ znázorňujú ako posadnutí slovami, vznášajúci sa v nekonečnom slovnom priestore oddelenom od skutočnosti.

Skutoční Tradiční Racionalisti však oddávna trvali na udržiavaní pevného spojenia so skúsenosťou:

> Ak si v učebnici chémie pozriete definíciu lítia, možno vám povie, že je to prvok, ktorého atómová hmotnosť je takmer presne 7. Ale ak autor rozmýšľal logickejšie, povie vám, že ak medzi minerálmi, ktoré sú sklovité, priesvitné, sivé alebo biele, veľmi pevné, krehké a nerozpustné, pohľadáte taký, ktorý dáva plameňu karmínový nádych, keď tento minerál rozotriete s vápnom alebo witheritom, a potom roztavíte, môžete ho čiastočne rozpustiť v kyseline soľnej; a keď tento roztok odparíte a zvyšok extrahujete kyselinou sírovou, a poriadne vyčistíte, dá sa bežnými metódami premeniť na chlorid, ktorý keď ho získate v pevnom stave, roztavíte a elektrolyzujete poltuctom silných batérií, vytvorí guľu ružovkastého strieborného kovu, ktorá pláva na benzíne; a tento materiál je vzorkou lítia.
>
> --Charles Sanders Peirce[165]

Toto je príklad „logickej mysle“, ako ju opistuje Tradičný Racionalista, a nie hollywoodsky scenárista.

Ale všimnite si: Peirce vám v skutočnosti *neukázal* kúsok lítia. Nemal kúsky lítia pribité vo svojej knihe. Namiesto toho vám dáva mapu pokladu – intenzionálne definovaný postup, ktorý vás po vykonaní privedie k extenzionálnemu príkladu lítia. Nie je to to isté, ako keby vám hodil kus lítia, ale tiež to nie je to isté ako povedať „atómová hmotnosť 7“. (Hoci keby ste mali *dostatočne ostré* oči, stačilo by vám povedať „tri protóny“ a možno by ste nejaké lítium zbadali...)

To je teda intenzionálna a extenzionálna *definícia*, čo je spôsob, ako povedať niekomu druhému, čo nejakým pojmom myslíte. Keď som predtým hovoril o „definíciách“, hovoril som o spôsobe, ako *komunikovať* pojmy – *povedať niekomu inému*, čo myslíte slovami „červená“, „tiger“, „človek“ alebo „lítium“. Teraz si povedzme o samotných pojmoch.

Skutočnou intenziou môjho pojmu „tiger“ by bol nervový vzor (v mojej spánkovej kôre), ktorý skúma prichádzajúci signál zo zrakovej kôry, aby určil, či to je alebo nie je tiger.

Skutočnou extenziou môjho pojmu „tiger“ je všetko, čo by som nazval tigrom.

---

165 Charles Sanders Peirce, *Collected Papers* (Harvard University Press, 1931).

Intenzionálne definície nezachytávajú celé intenzie; extenzionálne definície nezachytávajú celé extenzie. Ak ukážem na jedného tigra a poviem slovo „tiger“, komunikácia môže zlyhať, ak si pomyslia, že som tým myslel „nebezpečné zviera“ alebo „samec tigra“ alebo „žltá vec“. Podobne, ak poviem „nebezpečné žlto-čierno pásikavé zviera“ a neukážem pritom na nič, poslucháč si môže predstaviť veľkého sršňa.

Nedokážete slovami zachytiť všetky podrobnosti kognitívneho pojmu – ako existuje vo vašej hlave – ktorý vám umožňuje rozoznávať veci ako tigre alebo nie tigre. Je príliš veľký. A nemôžete ukázať prstom na všetky tigre, ktoré ste kedy videli, tobôž na všetko, čo *by ste* nazvali tigrom.

Najsilnejšie definície používajú krížovú paľbu intenzionálnej a extenzionálnej komunikácie, aby upresnili pojem. Ešte aj tak komunikujete iba *mapy* k pojmom, alebo návody na zostavenie pojmov – nekomunikujete *skutočné* kategórie, ako existujú vo vašej mysli alebo vo svete.

(Áno, pri dostatočnej tvorivosti dokážete vytvoriť výnimky z tohto pravidla, napríklad „Vety obsahujúce slovo ‚huragaloni‘, ktoré Eliezer Yudkowsky uverejnil 4. februára 2008.“ Práve som vám ukázal celú extenziu tohto pojmu. Ale mimo matematiky sú definície zvyčajne mapy k pokladom, nie samotné poklady.)

To je teda ďalší dôvod, prečo nemôžete „definovať slovo ako sa vám zachce“: Nemôžete priamo programovať pojmy do mozgu niekoho iného.

Ešte aj v rámci aristotelovskej paradigmy, kde predstierame, že definície sú skutočné pojmy, nemáte slobodu intenzie a extenzie *súčasne*. Predstavme si, že definujem Mars ako „Veľká červená kamenná guľa, s asi desatinou hmotnosti Zeme a o 50% ďalej od Slnka.“ Potom je samostatná úloha ukázať, že táto intenzionálna definícia zodpovedá nejakej konkrétnej extenzionálnej veci v mojej skúsenosti, a že vôbec zodpovedá nejakej skutočnej veci. Keby som namiesto toho povedal „Toto je Mars“ a ukázal na červené svetlo na nočnej oblohe, bolo by samostatnou úlohou ukázať, že toto extenzionálne svetlo zodpovedá nejakej konkrétnej intenzionálnej definícií, ktorú navrhnem – alebo nejakému intenzionálnemu názoru, ktorý mám – napríklad že „Mars je boh vojny“.

Ale väčšina činnosti mozgu pri používaní intenzií sa odohráva pod úrovňou vedomia. Neuvedomujeme si, že naša identifikácia červeného sveta ako „Mars“ je oddelená vec od našej slovnej definície „Mars je boh vojny.“ Bez ohľadu na to, akú intenzionálnu definíciu si vymyslím na opísanie Marsu, moja myseľ verí, že „Mars“ odkazuje na túto vec, a že je to štvrtá planéta slnečnej sústavy.

Ked vezmete do úvahy, ako ľudská myseľ naozaj v praxi funguje, z predstavy „môžem si definovať slovo, ako sa mi zachce“ sa rýchlo stane „môžem veriť, čomu sa mi zachce, ohľadom pevne danej množiny predmetov“ alebo „môžem presunúť predmet, ktorý sa mi zachce, do pevne daného testu členstva alebo von z neho“. Tak ako zvyčajne nedokážete slovami sprostredkovať celú intenziu pojmu, pretože je to zložitý neurónovy test členstva, nemôžete *ovládať* celú intenziu pojmu, pretože sa aplikuje podvedome. Preto je také obľúbené hádanie sa, že XYZ je pravdivé „z definície“. Keby sa zmeny definícií správali ako tie empirické nulové operácie, za ktoré ich pokladáme, nikto by sa neunúval argumentovať nimi. Ak však trochu zneužijete definície, premenia sa na čarovné paličky – samozrejme iba v argumentácii, nie v skutočnosti.

*→

157. Zhluky podobnosti

Kde bolo, tam bolo, filozofi z Platónovej Akadémie tvrdili, že najlepšia definícia človeka je „dvojnožec bez peria“. Diogenes zo Sinope, zvaný aj Diogenes Cynik, údajne pohotovo ukázal ošklbané kura a vyhlásil: „Tu je Platónov človek.“ Platónovci pohotovo zmenili svoju definíciu na „dvojnožec bez peria so širokými nechtami.“

Žiaden slovník a žiadna encyklopédia ešte nevymenovali všetky veci, ktoré majú ľudia spoločné. Máme červenú krv, po päť prstov na každej z dvoch rúk, kostnaté lebky, 23 párov chromozómov – ale to

→ http://lesswrong.com/lw/nh/extensions_and_intensions/

sa dá povedať aj o ďalších živočíšnych druhoch. Vyrábame zložité nástroje na výrobu zložitých nástrojov, používame syntaktický kombinačný jazyk, využívame kritické štiepne reakcie ako zdroj energie: tieto veci môžu poslúžiť na vybranie niektorých ľudí, ale nie všetkých – mnohí z nás nikdy nepostavili štiepny reaktor. Pomocou správnej sady potrebných a dostatočných génových sekvencií by ste mohli vybrať všetkých ľudí a iba ľudí – prinajmenšom teraz – ale stále by to bolo ďaleko od *všetkého*, čo majú ľudia spoločné.

Ale dokiaľ sa nenachádzate blízko ošklbaného kuraťa, povedať „Hľadaj dvojnožcov bez peria“ môže poslúžiť na vybranie niekoľkých tuctov konkrétnych vecí, ktoré sú ľudia, a nie domy, vázy, sendviče, mačky, farby, alebo matematické vety.

Keď už ste raz definíciu „dvojnožec bez peria“ spojili s nejakými *konkrétnymi* dvojnožcami bez peria, môžete sa pozrieť na túto skupinu a začať zhromažďovať niektoré *ďalšie* vlastnosti – okrem toho, že nemajú perie a majú dve nohy – ktoré majú tieto „dvojnožce bez peria“ spoločné. Tieto konkrétne dvojnožce bez peria, ktoré vidíte, zároveň používajú jazyk, budujú uložité nástroje, hovoria kombinačným jazykom so syntaxou, krvácajú na červeno, ak ich pichnete, zomrú, ak vypijú bolehlav.

Takto sa kategória „človek“ obohacuje a získava ďalšie a ďalšie vlastnosti; a keď Diogenes konečne prinesie svoje ošklbané kura, už nás neprekabáti: Toto ošklbané kura sa očividne nepodobá na ostatné „dvojnožce bez peria“.

(Keby bola aristotelovská logika dobrým modelom ľudskej psychológie, platónovci by sa boli pozreli na ošklbané kura a povedali: „Áno, toto je človek; čo si tým chcel povedať?“)

Ak *prvý* dvojnožec bez peria, ktorého vidíte, je ošklbané kura, potom si môžete pomyslieť, že slovné označenie „človek“ označuje ošklbané kura; takže ja môžem upraviť svoju mapu k pokladu, aby ukazovala na „dvojnožce bez peria so širokými nechtami“, a ak som múdry, ešte poviem: „Vidíš tam Diogena? To je človek, a ja som človek, a ty si človek; a tamten šimpanz nie je človek, ale celkom sa podobá.“

Úvodný náznak má používateľa iba priviesť k zhluku podobnosti – skupine vecí, ktoré majú nohé spoločné vlastnosti. Potom už úvodný náznak splnil svoj cieľ, a ja môžem pokračovať odovzdaním novej informácie „ľudia sú v súčasnosti smrteľní“, alebo čokoľvek iné chcem povedať o nás, dvojnožcoch bez peria.

Na slovník je najlepšie myslieť nie ako na knihu definícií aristotelovských tried, ale ako na knihu náznakov na priradenie slovných označení k zhlukom podobnosti, alebo na priradenie označení vlastnostiam, ktoré sú užitočné pri rozlišovaní zhlukov podobnosti.

*→

158. Typickosť a asymetrická podobnosť

Vtáky lietajú. Teda, okrem pštrosov. Ale čo je typickejší vták – vrabec alebo pštros?

Čo je typickejšia stolička: Kancelárska stolička, hojdacie kreslo, alebo vrece na sedenie?

Väčšina ľudí by povedala, že vrabec je typickejší vták a kancelárska stolička je typickejšia stolička. Kognitívni psychológovia, ktorí takéto veci študujú experimentálne, to robia pod nadpisom „efekt typickosti“ alebo „efekt prototypu“.[166] Ak napríklad požiadate pokusné osoby, aby stlačili tlačidlo vyjadrujúce „pravda“ alebo „nepravda“ v reakcii na výroky ako: „Vrabec je vták“ a „Tučniak je vták“, reakčné časy sú rýchlejšie pri ústrednejších príkladoch.[167] Miery typickosti dobre korelujú pri použití rôznych metód skúmania – jedna možnosť sú reakčné časy; ďalej môžete ľudí požiadať, aby priamo odhodnotili na škále od 1 do 10 ako dobre príklad (napríklad konkrétny vrabec) zapadá do kategórie (napríklad „vták“).

→ http://lesswrong.com/lw/nj/similarity_clusters/

166 Eleanor Rosch, „Principles of Categorization,“ in *Cognition and Categorization* [Poznávanie a kategorizácia], ed. Eleanor Rosch and Barbara B. Lloyd (Hillsdale, NJ: Lawrence Erlbaum, 1978).

167 George Lakoff, *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* [Ženy, oheň a nebezpečné veci: Čo nám hovoria kategórie o podstate myslenia] (Chicago: Chicago University Press, 1987).

Máme teda mentálnu mieru typickosti – ktorá by azda mohla fungovať ako heuristika – ale existuje aj zodpovedajúce skreslenie, pomocou ktorého ju môžeme upresniť?

Nuž, ktorý z týchto výrokov vám prípadá prirodzenejší: „98 je približne 100“ alebo „100 je približne 98“? Ak ste ako väčšina ľudí, prvý výrok vám pripadá zmysluplnejší.[168] Z podobných dôvodov ľudia, ktorých žiadate ohodnotiť, nakoľko sa Mexiko podobá na Spojené Štáty, dávajú konzistentne vyššie hodnotenie než ľudia, ktorých žiadate ohodnotiť, nakoľko sa Spojené Štáty podobajú na Mexiko.[169]

A ak vám aj toto pripadá neškodné, štúdia od Ripsa ukázala, že ľudia majú väčší sklon očakávať, že sa choroba na ostrove rozšíri z červienok na kačice, než z kačíc na červienky.[170] Toto síce nie je *logicky* nemožné, ale v pragmatickom zmysle, ak nejaký rozdiel odlišuje kačicu od červienky, a vďaka nemu by choroba mala menšiu šancu preniesť sa z kačice na červienku, musí to zároveň byť rozdiel medzi červienkou a kačicou, a preto by vďaka nemu mala choroba mať menšiu šancu preniesť sa z červienky na kačicu.

Áno, môžete si vymyslieť racionalizácie ako: „No, možno je tam viac živočíšnych druhov príbuzných červienkam, na ktoré by sa táto choroba mohla preniesť najprv, atď.,“ ale dávajte si pozor, aby ste sa príliš nesnažili racionalizovať pravdepodobnostné hodnotenia pokusných osôb, ktoré si ani neuvedomovali, že sa tam niečo porovnáva. A nezabudnite, že Mexiko sa viac podobá na Spojené Štáty než Spojené Štáty na Mexiko, a že 98 je bližšie k 100 než je 100 k 98. Jednoduchšie vysvetlenie je, že ľudia používajú (dokázanú) heuristiku podobnosti ako zástupcu pre pravdepodobnosť rozšírenia choroby, a že táto heuristika je (dokázateľne) asymetrická.

Kansas je nezvyčajne blízko k stredu Spojených Štátov, a Aljaška je nezvyčajne ďaleko od stredu Spojených Štátov; takže Kansas je pravdpeodobne bližšie k väčšine miest v USA a Aljaška je pravdepodobne ďalej. Z toho však nevyplýva, že Kansas je bližší k Aljaške než Aljaška ku Kansasu. Väčšina ľudí však rozmýšľa (metaforicky povedané) akoby blízkosť bola vnútornou vlastnosťou Kansasu, a vzdialenosť vnútornou vlastnosťou Aljašky; takže Kansas je vždy blízko, dokonca aj k Aljaške; a Aljaška je vždy ďaleko, dokonca aj od Kansasu.

Opäť raz vidíme, že aristotelovské chápanie kategórií – logických tried, ktorých členstvo je určené zbierkou vlastností, ktoré sú jednotlivo celkom nevyhnutné, a spolu celkom dostatočné – nie je dobrým modelom ľudskej kognitívnej psychológie. (Názor vedy sa za posledných 2350 rokov trochu zmenil? Kto by si to bol pomyslel?) Dokonca ani nerozmýšľame, akoby členstvo v skupine bola vlastnosť typu áno alebo nie: Tvrdenie o členstve v skupine môže byť viac alebo menej pravdivé. (Poznámka: To *nie je* to isté ako byť viac alebo menej pravdepodobné.)

Ďalší dôvod nepredstierať, že vy alebo ktokoľvek iný naozaj používa slová ako aristotelovské logické kategórie.

* →

159. Zhluková štruktúra priestoru vecí

Pojem „priestor konfigurácií“ je spôsob, ako *popisy* predmetov preložiť na *polohy* predmetov. Môže sa zdať, že modrá je „bližšie“ k modrozelenej než k červenej, ale o koľko bližšie? Na túto otázku je ťažké odpovedať iba pozeraním na farby. Ale pomôže vedieť, že (proporčné) farebné súradnice v RGB sú 0:0:5, 0:3:2 a 5:0:0. Ešte jasnejšie by to bolo po nakreslení na 3D graf.

Rovnakým spôsobom vidíte červienku ako červienku – hnedý chvost, červená hruď, štandardný červienkovský tvar, maximálna rýchlosť pri lete bez zaťaženia, DNA typická pre jej živočíšny druh a

168 Jerrold Sadock, „Truth and Approximations,“ [Pravda a aproximácie] *Papers from the Third Annual Meeting of the Berkeley Linguistics Society* (1977): 430–439.

169 Amos Tversky and Itamar Gati, „Studies of Similarity,“ [Štúdie podobnosti] in *Cognition and Categorization*, ed. Eleanor Rosch and Barbara Lloyd (Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1978), 79–98.

170 Lance J. Rips, „Inductive Judgments about Natural Categories,“ [Induktívne usudzovanie o prirodzených kategóriách] *Journal of Verbal Learning and Verbal Behavior* 14 (1975): 665–681.

→ http://lesswrong.com/lw/nk/typicality_and_asymmetrical_similarity/

jednotlivé alely. Alebo môžete vidieť červienku ako jeden bod v priestore konfigurácií, ktorého rozmery opisujú všetko, čo o červienke vieme, alebo by sme mohli vedieť.

Červienka je väčšia než vírus, a menšia než lietadlová loď – toto by mohla byť dimenzia „objem“. Podobne červienka váži viac než atóm vodíka, a menej než galaxia; to by mohla byť dimenzia „hmotnosť“. Rôzne červienky budú mať silnú koreláciu medzi „objemom“ a „hmotnosťou“, takže červienkové body budú v týchto dvoch dimenziách zoradené v pomerne lineárnej reťazi – ale korelácia nebude presná, takže budeme potrebovať dve samostatné dimenzie.

Toto je výhoda vnímania červienok ako bodov v priestore: Toto lineárne usporiadanie by ste tak ľahko nevideli, keby ste si červienky predstavovali iba ako malé milé tvory mávajúce krídlami.

DNA červienky je vysoko multidimenzionálna premenná, ale stále si ju môžeme predstaviť ako časť polohy červienky v priestore vecí – milióny štvorkových súradníc, jedna súradnica pre každú bázu DNA – alebo možno o čosi sofistikovanejší pohľad. Tvar červienky a jej farbu (odrážavosť povrchu) si možno tiež predstaviť ako súčasť polohy červienky v priestore vecí, aj keď to nie sú *jednotlivé* dimenzie.

Tak ako súradnice 0:0:5 obsahujú rovnakú informáciu ako farba modrá v HTML, nemali by sme strácať informáciu, keď vidíme červienky ako body v priestore. Veríme tomu istému výroku o hmotnosti červienky, či si predstavíme červienku vyváženú závažím 0,07 kg na opačnej miske váh, alebo si predstavíme bod červienka so súradnicou hmotnosť +70.

Môžeme si dokonca predstaviť priestor konfigurácií s jednou či viacerými dimenziami pre každú osobitnú vlastnosť predmetu, takže by *poloha* bodu v tomto priestore zodpovedala *všetkým* informáciám o danom skutočnom predmete. Bola by to pomerne redundantná reprezentácia – rozmery by zahŕňali hmotnosť, objem, aj hustotu.

Ak si myslíte, že je to extravagantné, kvantoví fyzici používajú *nekonečnorozmerné* priestory konfigurácií, kde jeden bod v priestore opisuje polohu každej častice vo vesmíre. V porovnaní s nimi sme pri našej predstave *priestoru vecí* pomerne konzervatívni – bod v priestore vecí opisuje iba jeden predmet, nie celý vesmír.

Ak si nie sme istí presnou hmotnosťou a objemom červienky, môžeme si predstaviť malý obláčik v priestore vecí, *objem neistoty*, v ktorom je táto červienka. Hustota toho oblaku je hustota nášho presvedčenia, že tá červienka má túto konkrétnu hmotnosť a objem. Ak ste si viac istí hustotou červienky než jej hmotnosťou a objemom, váš oblak pravdepodobnosti bude vysoko koncentrovaný v dimenzii hustoty, a koncentrovaný pozdĺž šikmej čiary v podpriestore hmotnosti a objemu. (Tento oblak je v skutočnosti plocha, kvôli vzťahu $V \times D = M$.)

„Lúčovité kategórie“ je pojem, ktorým kognitívni psychológovia opisujú nearistotelovské hranice slov. Centrálna „matka“ počne svoje dieťa, porodí ho, a vychová ho. Je darkyňa vajíčka, ktorá nikdy neuvidí svoje dieťa, matkou? Je to „genetická matka“. Čo žena, ktorej implantujú cudzie embryo a ona ho vynosí? Je to „náhradná matka“. A čo žena, ktorá vychová dieťa, ktoré nie je geneticky jej? Nuž, to je „adoptívna matka“. Aristotelovský sylogizmus by znel: „Ľudia majú desať prstov, Fred má deväť prstov, preto Fred nie je človek“, ale v skutočnosti rozmýšľame takto: „Ľudia majú desať prstov, Fred je človek, preto Fred je ‚deväťprstý človek‘.“

Túto lúčovitosť kategórií môžeme chápať v intenzionálnom zmysle, ako bolo opísané – vlastnosti, ktoré sú zvyčajne prítomné, ale občas môžu chýbať. Keby sme mysleli na intenziu slova „matka“, bolo by to ako rozložená žiara v priestore vecí, žiara, ktorej intenzita zodpovedá stupňu nakoľko daný objem v priestore vecí zodpovedá kategórii „matka“. Táto žiara je sústredená v strede genetiky a pôrodu a výchovy dieťaťa; objem darkyne vajíčka tiež svieti, ale menej jasne.

Alebo môžeme túto lúčovitosť chápať extenzionálne. Predstavme si, že by sme namapovali všetky vtáky v našom svete do priestoru vecí, používajúc metriku vzdialenosti, ktorá by čo najlepšie zodpovedala ľudskému vnímaniu podobnosti: Červienka sa viac podobá na inú červienku než niektorá z nich na holuba, ale holuby a červienky sa viac podobajú jedno na druhé než niektoré z nich na tučniaka, atď.

Potom by stred všetkej vtákovitosti bol husto osídlený mnohými susednými hustými zhlukmi, červienky a vrabce a kanáriky a holuby a mnohé dalšie druhy. Orly a sokoly a ďalšie veľké dravé vtáky by sídlili vo vedľajšom zhluku. Tučniaky by boli vo vzdialenejšom zhluku, podobne aj sliepky a pštrosy.

Výsledok by veru mohol vyzerať ako astronomický zhluk: veľa galaxií obiehajúcich okolo stredu, a niektoré pomimo.

Alebo by sme mohli zároveň myslieť aj na intenziu kognitívnej kategórie „vták“, aj na jej extenziu vo vtákoch v skutočnom svete: Ústredný zhluk červienok a vrabcov jasne žiari vysoko typickou vtákovitosťou; odľahlé zhluky pštrosov a tučniakov žiaria slabšie netypickou vtákovitosťou, a Abraham Lincoln je o niekoľko megaparsekov ďalej a nežiari vôbec.

Páči sa mi táto posledná predstava – žiariace body – pretože ako to vnímam, štruktúra kognitívnej intenzie nasledovala štruktúru extenzionálnych zhlukov. Najprv bola štruktúra v skutočnom svete, empirické rozloženie vtákov v priestore vecí; potom sme si ich pozorovaním vytvorili kategóriu, ktorej intenzionálna žiara približne pokrýva túto štruktúru.

To nám dáva ďalší pohľad na to, prečo slová nie sú aristotelovské triedy: empirická zhluková štruktúra skutočného sveta nie je taká kryštalická. Prirodzený zhluk, skupina vecí, ktoré sa navzájom vysoko podobajú, možno *nemá* množinu nevyhnutných a dostatočných podmienok – *nemá* množinu vlastností, ktoré všetci členovia majú a žiaden nečlen nemá.

Ale aj keď je nejaká kategória nenapraviteľne rozmazaná a hrboľatá, netreba panikáriť. Nenamietal by som, keby niekto povedal, že vtáky sú „operené lietajúce veci“. *Ale tučniaky nelietajú!* - no, v pohode. Zvyčajné pravidlo má výnimku; kvôli tomu sa svet nezrúti. Nemôžeme očakávať, že definície budú zakaždým presne zodpovedať empirickej štruktúre priestoru vecí, pretože mapa je menšia a omnoho menej zložitá než územie. Zmyslom definície „operené lietajúce veci“ je naviesť poslucháča na zhluk vtákov, nie dať mu úplný popis každého existujúceho vtáka až na molekulárnej úrovni.

Keď kreslíte hranicu okolo skupiny extenzionálnych bodov *empiricky* zoskupených v priestore vecí, možno nájdete aspoň jednu výnimku pre každé jednoduché intenzionálne pravidlo, ktoré vymyslíte.

Ale ak definícia v praxi funguje dosť dobre na to, aby ukázala na zamýšľaný empirický zhluk, namietanie voči nej môže byť oprávnene nazvané „hnidopišstvo“.

*→

160. Maskované otázky

Predstavte si, že máte nezvyčajnú prácu v nezvyčajnej továrni: vaša úloha je brať predmety z tajomného bežiaceho pásu a triediť ich do dvoch nádob. Keď prvýkrát prídete, Susan, seniorka triedička, vám vysvetlí, že tie modré predmety vajcovitého tvaru sa volajú „majcia“ a patria do „nádoby na majcia“, zatiaľ čo červené kocky sa volajú „čocky“ a patria do „nádoby na čocky“.

Keď začnete pracovať, všimnete si, že majcia a čocky sa odlišujú aj inými vecami okrem farby a tvaru. Majcia sú na povrchu chlpaté; čocky sú hladké. Majcia sa pri dotyku trochu prehýbajú; čocky sú tvrdé. Majcia sú matné; čocky majú mierne priesvitný povrch.

Krátko po začiatku práce natrafíte na majce, ktoré má nezvyčajne tmavý odtieň modrej – v skutočnosti sa po podrobnejšom prieskume táto farba ukáže ako fialová, napoly medzi červenou a modrou.

Hej, moment! Prečo tento predmet voláte „majce“? „Majce“ bolo pôvodne definované ako modré a vajcovitého tvaru – podmienka modrosti je v skutočnosti zahrnutá v samotnom názve „majce“. Tento predmet nie je modrý. Jedno z nevyhnutných kritérií tu chýba; mali by ste to nazvať „fialový predmet vajcovitého tvaru“, nie „majce“.

Lenže sa skrátka stalo, že okrem toho, že je fialový a vajcovitý, je tento predmet aj chlpatý, pružný a matný. Takže keď ste videli tento predmet, pomysleli ste si: „Aha, majce so zvláštnym odtieňom.“ Rozhodne to nie je čocka... však?

→ http://lesswrong.com/lw/nl/the_cluster_structure_of_thingspace/

Aj tak si nie ste celkom istí, čo máte urobiť. Takže zavoláte seniorku triedičku Susan.

„Ach áno, to je majce," povie Susan, „to môžeš dať do nádoby na majcia."

Začnete hádzať fialové majce do nádoby na majcia, ale na chvíľu sa zastavíte. „Susan," poviete, „ako *vieš*, že toto je majce?"

Susan na vás čudne pozrie. „Nie je to jasné? Tento predmet je síce fialový, ale stále je vajcovitý, chlpatý, pružný a matný, tak ako ostatné majcia. Musíš rátať s tým, že niektoré sú chybne sfarbené. Alebo to má byť jedna z filozofických hádaniek typu: ‚Ako vieš, že svet nebol stvorený pred piatimi minútami aj s našimi falošnými spomienkami?' Z filozofického hľadiska si nie som *absolútne istá*, že je to majce, ale pripadá mi to ako dobrý odhad."

„Nie, myslím tým..." Zastavíte sa, hľadáte slová. „Prečo je tu nádoba na majcia a nádoba na čocky? Čo je ten *rozdiel* medzi majcami a čockami?"

„Majcia sú modré a majú vajcový tvar, čocky sú červené a kockaté," povie Susan trpezlivo. „Bol si na štandardnom úvodnom zaškolení, však?"

„Prečo *vôbec* triedime majcia a čocky?"

„No... lebo inak by boli pomiešané?" povie Susan. „Pretože by nám nikto neplatil za to, aby sme tu celý deň iba vysedávali a *netriedili* majcia a čocky?"

„Kto to vlastne vymyslel, že prvý modrý vajcovitý predmet bude ‚majce' a ako na to došiel?"

Susan pokrčí plecami. „Predpokladám, že by si rovnako dobre mohol tie červené kockaté predmety nazvať ‚majcia' a tie modré vajcovité predmety ‚čocky', ale takto sa to ľahšie pamätá."

Chvíľu rozmýšľate. „Predstavme si, že by z bežiaceho pásu prišiel úplne domotaný predmet. Povedzme oranžový guľatý chpatý priehľadný predmet so zvíjajúcimi sa zelenými chápadlami. Ako by som zistil, či je to majce alebo čocka?"

„Wow, *taký* domotaný predmet ešte nikto nikdy nenašiel," povie Susan, „ale predpokladám, že by sme ho zobrali k triediacemu skeneru."

„Ako funguje ten triediaci skener?" vyzvedáte. „Rentgen? Magnetická rezonancia? Spektroskopia rýchlym neutrónovým prenosom?"

„Povedali mi, že funguje podľa Bayesovho pravidla, ale nerozumiem tomu, ako," povie Susan. „Ale rada to hovorím. Bayes Bayes Bayes Bayes Bayes."

„A čo ti *povie* ten triediaci skener?"

„Povie ti, či máš daný predmet položiť do nádoby na majcia alebo do nádoby na čocky. Preto sa to volá triediaci skener."

V tej chvíli vám dôjdu slová.

„Mimochodom," povie Susan mimochodom, „možno ťa zaujme, že majcia obsahujú malé nugety vanádovej rudy, a čocky obsahujú kúsky paládia, čo sú obe užitočné priemyselné suroviny."

„Susan, ty si čisté zlo."

„Ďakujem."

Teraz sa zdá, že sme objavili srdce a podstatu majcovitosti: majce je predmet, ktorý obsahuje nuget vanádovej rudy. Povrchné vlastnosti, ako modrá farba a chlpatosť, *neurčujú*, či je daný predmet majce;

povrchné vlastnosti sú dôležité iba tým, že vám pomôžu *usudzovať*, či daný predmet je majce, čiže či daný predmet obsahuje vanád.

Obsahovanie vanádu je potrebné a dostatočná definícia: všetky majcia obsahujú vanád a všetko, čo obsahuje vanád, je majce: „majce“ je iba skratka ako povedať „predmet obsahujúci vanád“. Správne?

Nie tak rýchlo, povie Susan: Zhruba 98 % majec obsahuje vanád, ale 2 % namiesto toho obsahuje paládium. Aby sme boli presní (pokračuje Susan), zhruba 98 % modrých vajcovitých chlpatých pružných matných predmetov obsahuje vanád. Pri nezvyčajných majciach to môže byť iné percento: 95 % fialových majec obsahuje vanád, 92 % tvrdých majec obsahuje vanád, atď.

Teraz si predstavte, že nájdete modrý vajcovitý chlpatý pružný matný predmet, obyčajné majce po každej viditeľnej stránke, a iba zo srandy ho vezmete k triediacemu skeneru, a skener povie „paládium“ - je to jedno z tých vzácnych 2 %. Je to majce?

Na prvý pohľad môžete odpovedať, že keď už tento predmet idete hodiť do nádoby na čocky, už ho rovno môžete nazvať „čocka“. Ukáže sa však, že takmer všetky majcia, keď vypnete svetlo, mierne svetielkujú v tme; zatiaľ čo takmer žiadne čocky v tme nesvetielkujú. Percento majec, ktoré svetielkuje v tme, nie je významne odlišné medzi modrými vajcovitými chlpatými matnými predmetmi, ktoré obsahujú paládium namiesto vanádu. Takže, ak chcete odhadnúť, či tento predmet svetielkuje ako majce, alebo zostane tmavý ako čocka, mali by ste odhadovať, že bude svetielkovať ako majce.

Takže, je tento predmet *naozaj* majce alebo čocka?

Na jednej strane, bez ohľadu na to, čo ďalšie sa dozviete, hodíte tento predmet do nádoby na čocky. Na druhej strane, ak sú nejaké neznáme vlastnosti predmetu, ktoré potrebujete odvodiť, budete ich odvodzovať, akoby tento predmet bol majce, nie čocka – zaradíte ho do zhluku podobnosti modrých vajcovitých chlpatých pružných matných vecí, a nie do zhluku podobnosti červených kockatých hladkých priesvitných vecí.

Veta „Je tento predmet majce?“ môže za rôznych okolností predstavovať rôzne otázky.

Keby nevyjadrovala *nejakú* otázku, nemali by ste dôvod sa o to starať.

Je ateizmus „náboženstvo“? Je transhumanizmus „sekta“? Ľudia, ktorí argumentujú, že ateizmus je náboženstvo „pretože vyjadruje názor na Boha“ sa naozaj snažia povedať (aspoň si myslím), že rozhodovacie procesy používané ateistami sú na rovnakej úrovni ako rozhodovacie procesy používané v náboženstve, alebo že ateizmus nie je o nič bezpečnejší než náboženstvo z hľadiska pravdepodobnosti podnecovania násilia, atď... O čo tu naozaj ide, je ateistovo tvrdenie o podstatnom rozdiele a nadradenosti voči náboženstvu, čo sa veriaci človek snaží odmietnuť tým, že popiera rozdiel namiesto nadradenosti(!).

Ale to nie je tá apriori nerozumná časť: Apriori nerozumná časť je tá, kde v priebehu hádky niekto vytiahne slovník a vyhľadá definície pre „ateizmus“ a „náboženstvo“. (A áno, je to rovnako hlúpe, či to urobí ateista alebo veriaci.) Ako by slovník *vôbec* mohol rozhodnúť, či je empirický zhluk ateistov podstatne odlišný od empirického zhluku teológov? Ako sa môže skutočnosť meniť podľa významu slova? Body v priestore vecí sa nepohybujú, keď okolo nich prekreslíme hranicu.

Lenže ľudia si často *neuvedomujú*, že ich hádka o tom, kde nakresliť hranicu definície, je v skutočnosti debatou o tom, či odvodzovať vlastnosti, ktoré má spoločné väčšina vecí v empirickou zhluku...

Preto hovorím, „maskované otázky“.

*→

161. Neurónové kategórie

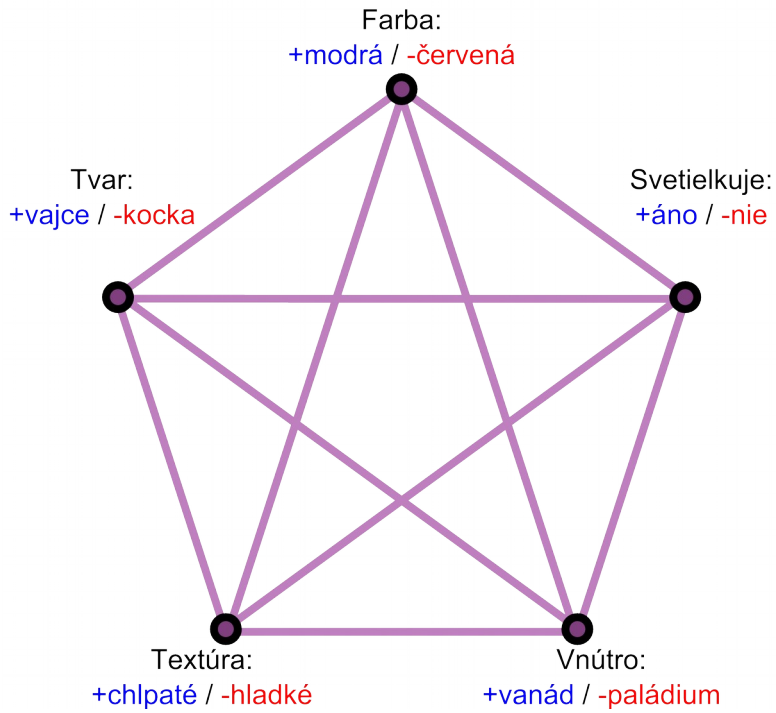
V maskovaných otázkach som hovoril o úlohe triediť „majcia“ a „čocky“. Typické majce je modré, vajcovité, chlpaté, pružné, matné, svetielkuje v tme, a obsahuje vanád. Typická čocka je červená, kockatá, hladká, tvrdá, priesvitná, nesvetielkuje, a obsahuje paládium. Kvôli jednoduchosti zabudnime na

→ http://lesswrong.com/lw/nm/disguised_queries/

vlastnosti pružnosť/tvrdosť a matnosť/priesvitnosť. To nám v priestore vecí necháva päť dimenzií: Farba, tvar, textúra, svietivosť a vnútro.

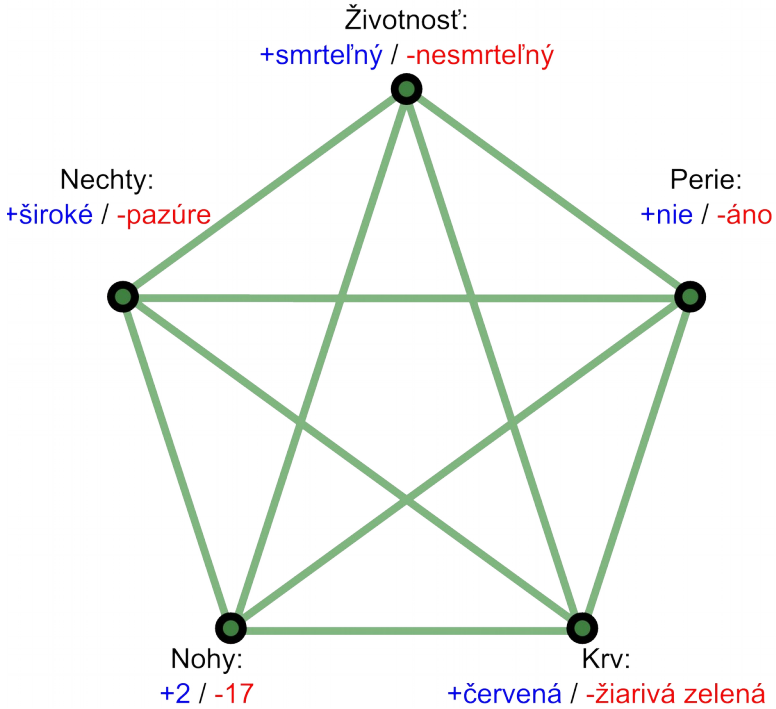
Predstavme si, že chcem vytvoriť umelú neurónovú sieť (UNS) na predpovedanie nepozorovaných vlastností majec na základe ich pozorovaných vlastností. A predstavme si, že som ohľadom UNS veľmi naivný: prečítal som si vzrušujúce populárnovedecké knihy o tom, že neurónové siete sú distribuované, emergentné a paralelné, *rovnako ako ľudský mozog!!* neviem však odvodiť diferenciálne rovnice pre pokles gradientu v nerekurentnej viacvrstvovej sieti so sigmoidnými jednotkami (čo je v skutočnosti omnoho ľahšie než to znie).

Potom by som mohol navrhnúť neurónovú sieť, ktorá bude vyzerať ako obrázok 161.1.



Obrázok 161.1: Sieť 1

Sieť 1 je na triedenie majec a čociek. Ale keďže „majce“ je neznámy a umelý pojem, pridal som aj podobnú sieť 1b na obrázku 161.2 na triedenie ľudí a Vesmírnych Príšer na základe údajov od Aristotela („všetci ľudia sú smrteľní“) a Platónovej akadémie („dvojnožec bez krídel so širokými nechtami“).



Obrázok 161.2: Sieť 1b

Neurónová sieť potrebuje pravidlo na učenie. Samozrejmý nápad je, že keď sú dva uzly často aktívne zároveň, mali by sme posilniť spojenie medzi nimi – to je jedno z prvých pravidiel, ktoré boli kedy na výcvik neurónovej siete navrhnuté, známe ako Hebbovo pravidlo.

Keby ste teda často videli veci, ktoré boli zároveň modré a chlpaté – čím zároveň aktivovali uzol „farba“ v stave +1 a uzol „textúra“ v stave +1 – posilnilo by sa spojenie medzi farbou a textúrou, takže by kladné farby aktivovali kladné textúry a naopak. Keby ste videli veci, ktoré sú modré a vajcovité a obsahujú vanád, posilnilo by to vzájomné kladné spojenia medzi farbou a tvarom a vnútrom.

Povedzme, že už ste videli hromady majec a čociek na bežiacom páse. Ale teraz vidíte niečo, čo je chlpaté, vajcovité, a – ach! - červenofialové (čo budeme modelovať ako úroveň aktivácie „farby“ -2/3). Netestovali ste svietivosť ani vnútro. Čo budete predpovedať, čo budete predpovedať?

Stane sa to, že úrovne aktivácie v sieti 1 budú trochu poskakovať. Kladná aktivácia prichádza z tvaru do svietivosti, záporná aktivácia prichádza z farby do vnútra, záporná aktivácia prichádza z vnútra do svietivosti... Samozrejme všetky tieto správy sa odovzdávajú *paralelne!!* a *asynchrónne!!* rovnako ako v ľudskom mozgu...

Nakoniec sa sieť 1 ustáli v stabilnom stave, ktorý má vysokú kladnú aktiváciu pre „svietivosť“ a „vnútro“. Môžeme teda povedať, že sieť „očakáva“ (aj keď ešte nevidela), že predmet bude svietiť v tme a bude obsahovať vanád.

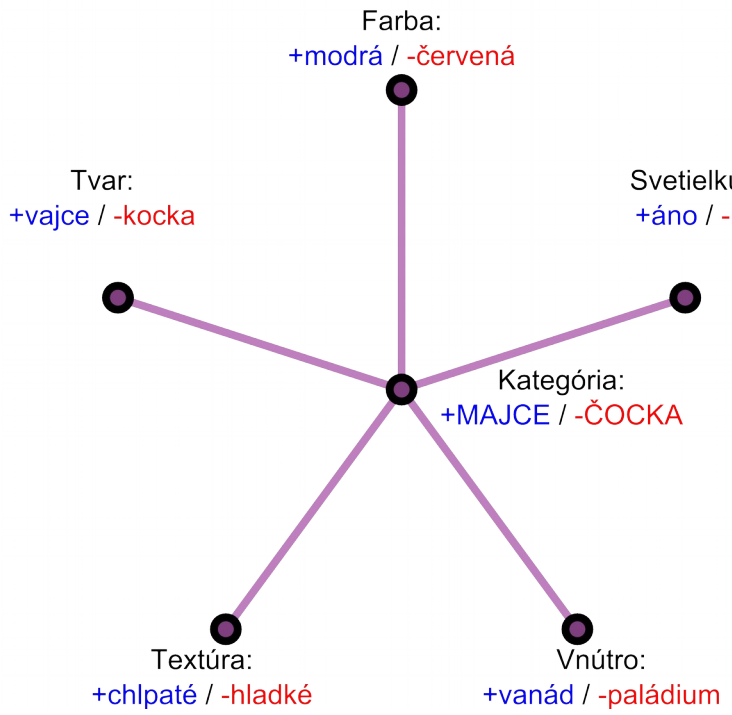
A hľa, sieť 1 vykazuje toto správanie aj keď nemá žiaden explicitný uzol hovoriaci, či daný objekt je majce alebo nie. Tento úsudok je *implicitný v celej sieti!!* Majcovitosť je *atraktor!!* ktorý vzniká ako výsledok *emergentného správania!!* z *distribuovaných!!* pravidiel učenia.

V skutočnom živote však tento druh návrhu siete – nech znie akokoľvek moderne – naráža na *všemožné* problémy. Rekurentné siete sa nemusia hneď ustáliť: Môžu kmitať alebo vykazovať chaotické správanie, alebo im len môže veľmi dlho trvať, než sa ustália. To je Zlá Vec, ak vidíte niečo veľké a žlté a pásikavé, a musíte počkať päť minút, kým sa vaša distribuovaná neurónová sieť ustáli v atraktore „tiger“. Môže to byť asynchrónne a paralelné, ale nie je to v reálnom čase.

A sú tu ďalšie problémy, ako napríklad dvojité započítavanie indície, keď sa správy odrážajú tam a naspäť: Ak máte podozrenie, že nejaký predmet svetielkuje v tme, vaše podozrenie aktivuje dojem, že ten predmet obsahuje vanád, čo opäť aktivuje dojem, že ten predmet svetielkuje v tme.

Navyše, keby ste sa pokúsili rozšíriť dizajn siete 1, vyžadovalo by si to $O(N^2)$ spojení, kde N je celkový počet pozorovateľných vlastností.

Čo by teda mohol byť realistickejší dizajn neurónovej siete?



Obrázok 161.3: Sieť 2

V sieti 2 na obrázku 161.3 sa vlna aktivácie zo všetkých zapojených (pozorovaných) uzlov spojí na centrálnom uzle, a potom sa odrazí späť do všetkých nezapojených (nepozorovaných) uzlov. Čo znamená, že môžeme odpoveď vypočítať na dva kroky, namiesto čakania, až sa sieť ustáli – dôležitá požiadavka v biológii, kde neuróny fungujú iba rýchlosťou 20 Hz. A architektúra siete rastie ako O(N) namiesto $O(N^2)$.

Uznávam, že niektoré veci si všimnete ľahšie pomocou prvej architektúry siete než pomocou druhej. Sieť 1 má priame spojenie medzi každými dvoma uzlami. Ak teda červené predmety *nikdy* nesvietia v tme, ale červené chlpaté predmety zvyčajne majú všetky ostatné vlastnosti majec, ako vajcovitý tvar a vanád, sieť 1 toto dokáže ľahko reprezentovať: potrebuje na to iba veľmi silné priame záporné spojenie z farby na svietivosť, ale omnoho silnejšie kladné spojenia z textúry na všetky zvyšné uzly okrem svietivosti.

Navyše toto nie je „špeciálna výnimka“ zo všeobecného pravidla, že majcia svietia – pamätajte, v sieti 1 nie je žiadna jednotka, ktorá reprezentuje majcovitosť; majcovitosť sa vynára ako atraktor distribuovanej siete.

Takže áno, za tých N^2 spojení sme niečo dostali. Ale nie veľa. Sieť 1 nie je *omnoho* užitočnejšia pri väčšine problémov v skutočnom svete, kde málokedy nájdete zviera zaseknuté v polovici medzi tým, či je mačka alebo pes.

(Existujú aj fakty, ktoré nedokážete ľahko reprezentovať *ani* v sieti 1 *ani* v sieti 2. Povedzme, že belasá farba a guľový tvar, keď sa vyskytujú pohromade, vždy naznačujú prítomnosť paládia; ale keď sa vyskytujú oddelene, jedno bez druhého, vždy veľmi silne naznačujú vanád. Toto sa v oboch architektúrach veľmi ťažko reprezentuje bez dodatočných uzlov. Aj sieť 1 aj sieť 2 stelesňujú implicitné predpoklady o tom, v akom druhu prostredia budú pravdepodobne existovať; schopnosť vyčítať toto je to, čo oddeľuje dospelých od kojencov v oblasti strojového učenia.)

Nemýľte sa: Ani sieť 1 ani sieť 2 nie sú biologicky realistické. *Ale* stále vyzerá ako férový odhad, že akokoľvek mozog naozaj funguje, v nejakom zmysle sa to viac podobá na sieť 2 než na sieť 1. Rýchle, lacné, rozšíriteľné, dokáže dobre rozoznať psy a mačky: prirodzený výber ide za takýmito vecami ako keď voda tečie nadol po teréne spôsobilosti.

Vyzerá to ako dostatočne jednoduchá úloha triediť predmety ako majcia a čocky, hádzať ich do príslušnej nádoby. Ale všimli by ste si, keby belasé predmety nikdy nesvetielkovali v tme?

Možno, keby vám niekto priniesol dvadsať predmetov, ktoré by mali spoločné iba to, že sú belasé, a potom by zhasol svetlo a žiaden z týchto predmetov by nesvetielkoval. Inými slovami, keby vám to obúchal o hlavu. Možno tým, že by vám predložil všetky tieto belasé predmety pohromade, by si váš mozog vytvoril novú podkategóriu, a dokázal by si všimnúť vlastnosť „nesvieti“ v tejto podkategórii. Ale pravdepodobne by ste si to nevšimli, keby tie belasé predmety boli roztrúsené medzi stovkou iných majec a čociek. Nebolo by *ľahké* ani *intuitívne* si to všimnúť takým spôsobom, ako je ľahké a intuitívne rozlišovať psy a mačky.

Alebo: „Sokrates je človek, všetci ľudia sú smrteľní, preto Sokrates je smrteľný.“ Ako vedel Aristoteles, že Sokrates je človek? Nuž, Sokrates nemal perie, mal široké nechty, chodil vzpriamene, hovoril po grécky, a celkovo mal tvar ako človek, aj sa tak správal. Mozog sa preto rozhodol, raz a navždy, že Sokrates je človek; a odvtedy odvodzuje, že Sokrates je smrteľný tak ako všetci dovtedy pozorovaní ľudia. Nezdá sa príliš ľahké ani intuitívne pýtať sa, nakoľko je nosenie šiat, na rozdiel od používania jazyka, spojené so smrteľnosťou. Iba „veci, ktoré nosia šaty a používajú jazyk, sú ľudia“ a „ľudia sú smrteľní“.

Existujú skreslenia spojené so snahou zatrieďovať veci do kategórií raz a navždy? Samozrejme, že áno. Pozrite si napríklad Sektárske antisektárstvo.

*→

→ http://lesswrong.com/lw/nn/neural_categories/

162. Ako sa algoritmus cíti zvnútra

„Keď spadne strom v lese a nikto to nepočuje, vydá pritom zvuk?“ Pamätám si, že som videl skutočnú hádku, ktorá začala na túto tému – celkom naivný argument, ktorý zďaleka nešiel okolo Berkeleyovského subjektivizmu. Iba:

„Vydá pritom zvuk, rovnako ako hocijaký iný padajúci strom!“

„Ale ako to môže byť zvuk, keď to nikto nepočuje?“

Štandardný racionalistický pohľad by mohol byť, že prvá osoba hovorí akoby „zvuk“ znamenalo zvukové vibrácie vo vzduchu; druhá osoba hovorí akoby „zvuk“ znamenalo sluchový vnem v mozgu. Ak sa opýtate: „Budú tam vzduchové vibrácie?“ alebo „Budú tam sluchové vnemy?“, odpoveď je hneď jasná. A tak je v skutočnosti celé hádka o definícii slova „zvuk“.

Myslím si, že táto štandardná analýza je v podstate správna. Prijmime to teda ako predpoklad a opýtajme sa: Prečo sa ľudia hádajú týmto spôsobom? Aká psychológia je za tým?

Kľúčovou myšlienkou programu heuristík a skreslení je, že chyby často odhaľujú o myslení viac než správne odpovede. Dostať sa do zapálenej debaty o tom, či, keď strom padne v opustenom lese, vydá zvuk, sa tradične považuje za chybu.

Aký druh dizajnu mysle teda zodpovedá takejto chybe?

V maskovaných otázkach som zaviedol úlohu rozlošovania majec/čociek, pri ktorej seriorka testerka Susan vysvetí, že vaša úloha je triediť predmety prichádzajúce na bežiacom páse, modré vajcia čiže „majcia“ ukladať do jednej nádoby, a červené kocky čiže „čocky“ do druhej nádoby. Ako sa ukáže, je to preto, lebo majcia obsahujú malé nugety vanádovej rudy, a čocky obsahujú malé kúsky paládia, čo je oboje priemyselne užitočné.

Akurát, že zhruba 2 % modrých vajcovitých predmetov namiesto toho obsahuje paládium. Ak teda nájdete modrú vajcovitú vec, ktorá obsahuje paládium, mali by ste ju skôr nazvať „čocka“? Keď ju idete dať do nádoby na čocky – prečo ju nenazvať „čocka“?

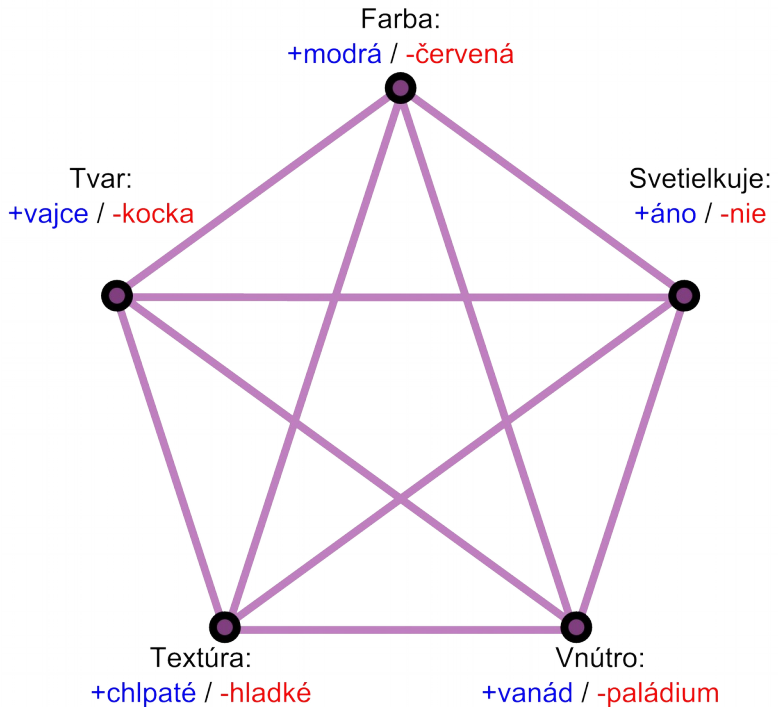
Lenže keď zhasnete svetlo, takmer všetky majcia v tme slabo svetielkujú. A modré vajcovité predmety obsahujúce paládium majú rovnakú pravdepodobnosť, že budú v tme svetielkovať, ako ľubovoľný iný modrý vajcovitý predmet.

Ak teda nájdete modrý vajcovitý predmet, ktorý obsahuje paládium a opýtate sa: „Je to majce?“, odpoveď záleží na tom, čo chcete s touto odpoveďou robiť. Ak sa pýtate: „Do ktorej nádoby tento predmet patrí?“, potom konajte, akoby tento predmet bola čocka. Ale ak sa pýtate: „Keď zhasnem svetlo, bude to svietiť?“, potom predpovedajte, akoby tento predmet bolo majce. V prvom prípade sa za vetou: „Je to majce?“ skrýva otázka: „Do ktorej nádoby to mám dať?“. V druhom prípade sa za vetou: „Je to majce?“ skrýva otázka: „Bude to v tme svietiť?“

Teraz si predstavte, že máte predmet, ktorý je modrý a vajcovitý a obsahuje paládium; a vy už ste pozorovali, že je chlpatý, pružný, matný a svetielkuje v tme.

Toto odpovedá na *všetky* otázky, pozoruje každú známu pozorovateľnú veličinu. Nezostalo nič, *čo* by maskovaná otázka mohla znamenať.

Prečo by niekto mohol cítiť nutkanie pokračovať v hádke, či tento objekt je *naozaj* majce?

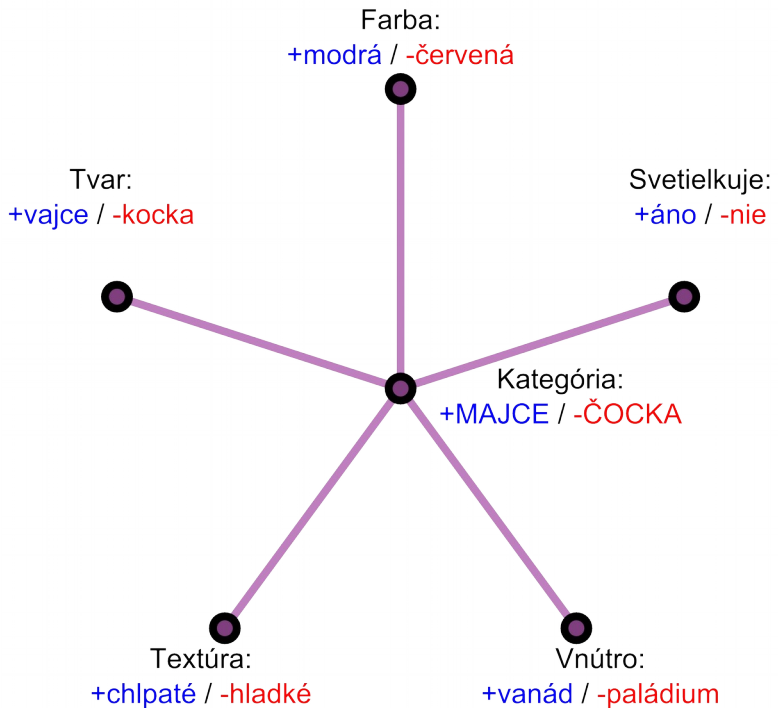


Obrázok 162.1: Sieť 1

Tieto diagramy z kapitoly Neurónové kategórie zobrazujú dve rôzne neurónové siete, ktoré mohli byť použité na odpovedanie otázok o majciach a čockách. Sieť 1 (obrázok 162.1) má veľa nevýhod – napríklad možné kmitavé/chaotické správanie alebo vyžadovanie $O(N^2)$ spojení – ale štruktúra siete 1 má oproti sieti 2 jednu veľkú výhodu: Každá jednotka v sieti zodpovedá testovateľnej otázke. Keď pozorujete všetky pozorovateľné veličiny, nezostali vám žiadne jednotky v sieti navyše.

Sieť 2 (obrázok 162.2) je však omnoho lepší kandidát na niečo, čo sa hmlisto podobá na fungovanie ľudského mozgu: Je rýchla, lacná, škálovateľná – a v strede má jednu voľne visiacu jednotku, ktorej aktivácia sa stále môže meniť, ešte aj keď sme pozorovali každý jeden z okolitých uzlov.

Čo inými slovami znamená, že aj potom, čo viete, či je nejaký predmet modrý alebo červený, vajcovitý alebo kockatý, chlpatý alebo hladký, svietivý alebo tmavý, a či obsahuje vanád alebo paládium, stále máme *pocit*, že tu zostala ešte jedna nezodpovedaná otázka: *Ale je to naozaj majce?*



Obrázok 162.2: Sieť 2

V našej každodennej skúsenosti sa zvukové vibrácie a sluchové vnemy vyskytujú spolu. Strom padajúci v opustenom lese však toto spojenie oddeľuje. A ešte aj keď viete, že padajúci strom vytvára zvukové vibrácie, ale nie sluchové vnemy, stále *máte pocit*, že zostala jedna nezodpovedaná otázka: *Vydal pritom zvuk?*

Vieme, kde je Pluto, a kam smeruje; poznáme tvar Pluta aj hmotnosť Pluta – ale je to planéta?

Teraz si zapamätajte: Keď sa pozriete na sieť 2, ako som ju tu nakreslil, vidíte tento algoritmus zvonku. Ľudia si nekladú otázku: „Tak mala by tá centrálna jednotka signalizovať, alebo nie?“ rovnako ako nemyslíte na: „Mal by neurón číslo 12 234 320 242 v mojej zrakovej kôre vyslať signál alebo nie?“

Vyžaduje si to vedomé úsilie predstaviť si váš mozog zvonku – a ani potom ešte nevidíte svoj skutočný mozog; predstavujete si, čo si *myslíte*, že tam je, v lepšom prípade na základe vedy, ale aj tak, nemáte k štruktúram neurónových sietí žiaden priamy prístup na introspekciu. Preto starovekí Gréci nevynašli výpočtovú neurovedu.

Keď sa pozriete na sieť 2, vidíte ju *zvonku*; ale ako sa táto neurónová sieť cíti *zvnútra*, ak vy sám *ste* mozog, v ktorom beží tento algoritmus, to je že aj keď poznáte každú vlastnosť daného predmetu, stále si kladiete otázku: „Ale je to teda majce, alebo nie?“

Toto je veľká trhlina na prekročenie, a videl som, ako to ľudí zastavilo. Pretože svoje vlastné intuície nevidíme ako „intuície“; vidíme ich ako svet. Keď sa pozriete na zelenú šálku, nepredstavujete si sami seba, ako vidíte obrázok rekonštruovaný vo vašej zrakovej kôre – hoci to *je* to, čo vidíte – jednoducho vidíte zelenú šálku. Pomyslíte si: „Aha, táto šálka je zelená“ a nie: „Obrázok tejto šálky v mojej zrakovej kôre je zelený.“

A rovnako, keď sa ľudia hádajú o tom, či padajúci strom vydá zvuk, alebo či je Pluto planéta, nevidia sami seba ako hádajúcich sa, či má v ich neurónovej sieti aktívna nejaká kategorizácia. Im to pripadá tak, že ten strom buď vydá zvuk, alebo nie.

Vieme, kde je Pluto a kam ide; poznáme tvar a hmotnosť Pluta – ale je to planéta? Áno, sú ľudia, ktorí povedia, že toto je hádka ohľadom definícií – ale ešte aj toto je uhol pohľadu siete 2, pretože sa hádate o tom, ako by táto centrálna jednotka mala byť nastavená. Keby ste boli myseľ zostavená podľa modelu siete 1, nepovedali by ste: „To závisí od toho, ako si definujeme ‚planétu‘“; povedali by ste iba: „Ak poznáme obežnú dráhu a tvar a hmotnosť Pluta, už sa nie je čo pýtať.“ Alebo skôr, tak by ste to *cítili* – mali by ste *pocit*, že už nezostala žiadna otázka – keby ste boli myseľ zostavená podľa vzoru siete 1.

Skôr než začnete skúmať svoje intuície, musíte si uvedomiť, že to, na čo sa pozerá vaša myseľ, *je* intuícia – nejaký poznávací algoritmus, aký sa dá vidieť zvonku – a nie priame vnímanie toho, Ako Veci Naozaj Sú.

Myslím si, že ľudia lipnú na svojich intuíciách ani nie preto, že by verili, že ich poznávacie algoritmy sú dokonale spoľahlivé, ale pretože nedokážu vidieť svoje intuície *ako to, aké to je, keď sa poznávací algoritmus pozerá zvnútra.*

A preto všetko, čo sa pokúšate povedať o tom, ako sa prirodzené poznávacie algoritmy mýlia, postavia do kontrastu so svojím priamym vnímaním Ako Veci Naozaj Sú – a zahodia to ako očividne nesprávne.

* →

163. Debaty o definíciách

Sledoval som už viac než jeden rozhovor – dokonca aj rozhovory, ktoré údajne mali byť o kognitívnej vede – vedúce k debate o definíciách. Ak si vezmeme klasický príklad: „Ak strom padne v lese, a nikto to nepočuje, vydá pritom zvuk?“, debata často nadobudne takéto smerovanie:

Ak strom padne v lese, a nikto to nepočuje, vydá pritom zvuk?

→ http://lesswrong.com/lw/no/how_an_algorithm_feels_from_inside/

Albert: „Samozrejme, že áno. Čo je to za hlúpu otázku? Vždy, keď som počul ako strom padol, vydal zvuk, takže si myslím, že aj iné padajúce stromy vydávajú zvuky. Neverím tomu, že sa svet mení, keď sa naň nepozerám.“

Barry: „Počkaj chvíľu. Ak to nikto nepočuje, ako to môže byť zvuk?“

V tomto príklade sa Barry háda s Albertom, pretože majú naozaj odlišnú intuíciu o tom, čo je podstatou zvuku. Ale existuje viac než jeden spôsob, ako môže začať Štandardná Debata. Barry môže mať motív odmietnuť Albertov záver. Alebo Barry môže byť skeptik, ktorý keď počuje Albertov argument, reflexívne prešetril, či neobsahuje logické chyby; a potom, keď našiel protiargument, automaticky ho prijal bez použitia druhej vrstvy hľadania proti-protiargumentu; čím sa sám vyargumentoval do opačného postoja. Toto si nevyžaduje, aby mal Barry opačnú *pôvodnú* intuíciu – intuíciu, ktorú Barry mohol mať, keby sme sa ho opýtali skôr než Albert prehovoril – inú ako Albert.

Nuž, ak aj Barry pôvodne nemal odlišnú intuíciu, teraz ju už určite má.

Albert: „Čo tým myslíš, že nie je zvuk? Korene stromu prasknú, kmeň sa zrúti a narazí na zem. To vytvorí chvenie, ktoré sa šíri zemou a vzduchom. Tam ide energia padajúceho stromu, do tepla a zvuku. Hovoríš, že keď ľudia odídu z lesa, strom porušuje zákon zachovania energie?“

Barry: „Ale nikto nič nepočuje. Ak v lese nie sú žiadni ľudia, alebo, za účelom argumentu, nič iné so zložitou nervovou sústavou schopnou ‚počuť‘, potom nikto nepočuje žiaden zvuk.“

Albert a Barry zhromažďujú argumenty, ktoré *cítia* ako podporu pre ich jednotlivé postoje, opisujú čoraz detailnejšie, čo spôsobilo, že ich detektor „zvuku“ vyslal signál alebo zostal ticho. Ale zatiaľ sa rozhovor stále sústredí na les, nie na definície. A všimnite si, že si v skutočnosti neodporujú o ničom, čo sa v lese stalo.

Albert: „Toto je tá najhlúpejšia hádka, akú som kedy zažil. Si totálny ťulpas.“

Barry: „Áno? No ty zase vyzeráš ako mechom udretý.“

Keď si vymenili urážky, môžu sa obe strany stiahnuť a nestratiť pritom tvár. Z technického hľadiska to nie je súčasťou *debaty*, ako takéto veci chápu racionalisti; ale je to dôležitá časť Štandardnej Debaty, takže to spomínam tiež.

Albert: „Strom vydáva akustické chvenie. Z definície je to zvuk.“

Barry: „Nikto nič nepočuje. Z definície to nie je zvuk.“

Ohnisko hádky sa začína posúvať k definíciám. Kedykoľvek ste v pokušení povedať slová „z definície“ v hádke, ktorá nie je doslova o čistej matematike, pamätajte, že čokoľvek, čo je pravda „z definície“, je pravdivé vo všetkých možných svetoch, takže pozorovanie pravdivosti toho vám nemôže obmedziť, v *ktorom* svete žijete.

Albert: „Mikrofón môjho počítača dokáže nahrať zvuk, aj keď nie je nablízko nikto, kto by ho počul, uložiť ho do súboru, a volá sa to ‚zvukový súbor‘. A v tom súbore sú uložené vzory chvenia vzduchu, nie vzory neurónových signálov v niekoho mozgu. ‚Zvuk‘ znamená vzor chvenia.“

Albert vyloží argument, ktorý *cíti* ako podporu pre to, že slovo „zvuk“ *má istý konkrétny význam*. To je iný druh otázky, než či sa v lese vyskytujú akustické vibrácie – ale tento posun zvyčajne prejde nepovšimnutý.

Barry: „Vážne? No pozrime sa, či s tebou slovník súhlasí.“

Je veľa vecí, ktoré by ma mohli zaujímať v prípade padajúceho stromu. Mohol by som ísť do lesa a pozrieť sa na stromy, alebo sa naučiť, ako odvodiť zvukovú rovnicu zmien tlaku vzduchu, alebo preskúmať anatómiu ucha, alebo študovať neuroanatómiu sluchovej kôry. Namiesto robenia hocičoho z uvedeného by som si asi mal prečítať slovník. Prečo? Sú redaktori slovníka odborníci na botaniku,

odborníci na fyziku, odborníci na neurovedu? Pozrieť si encyklopédiu, to môže dávať zmysel, ale prečo *slovník*?

Albert: „Ha! Definícia 2c v Merriamovi-Websterovi: ‚Zvuk: Mechanická šíriaca sa energia prenášaná pozdĺžnymi tlakovými vlnami v hmotnom prostredí (napríklad vzduch).‘“

Barry: „Ha! Definícia 2b v Merriamovi-Websterovi: ‚Zvuk: Vnem vnímaný sluchovým orgánom.‘“

Albert a Barry, zborovo: „Prekliaty slovník! Vôbec nepomáha!“

Redaktori slovníkov sú historikmi používania, nie zákonodarcami jazyka. Redaktori slovníkov hľadajú aktuálne používané slová, potom ich napíšu spolu s tým (malou časťou toho), čo nimi ľudia označujú. Ak existuje viac ako jeden spôsob použitia, redaktori napíšu viac ako jednu definíciu.

Albert: „Pozri, predpokladajme, že by som nechal v lese mikrofón a nahral by som vzor akustických vibrácií padajúceho stromu. Keby som to niekomu pustil, nazval by to ‚zvuk‘! To je bežný význam tohto slova! Nevymýšľaj si svoje vlastné pochabé definície!“

Barry: „Po prvé, môžem definovať slovo tak, ako sa mi chce, pokiaľ ho používam konzistentne. Po druhé, použil som zmysel, ktorý bol v slovníku. Po tretie, kto ti dal právo rozhodovať, čo je a čo nie je bežný význam?“

V tejto Štandardnej Debate je veľa chýb rozumnosti. Niektoré z nich už som spomínal, ďalšie ešte len spomeniem; podobne aj s liekmi na ne.

Teraz chcem iba poukázať – trúchlivým spôsobom – že Albert a Barry sa napohľad zhodujú v prakticky každej otázke ohľadom toho, čo sa v lese *naozaj* deje, a predsa sa nezdá, že by to vytváralo nejaký pocit súhlasu.

Hádka o definíciách je ako chodník v parku; ľudia by nešli po tom chodníku, keby hneď od začiatku videli, kam vedie. Keby ste sa opýtali Alberta (Barryho), prečo sa ešte háda, pravdepodobne by povedal niečo ako: „Barry (Albert) sa pokúša zaviesť svoju vlastnú definíciu ‚zvuku‘, lotor prešibaný, aby podporil svoje smiešne tvrdenie; a ja tu obraňujem štandardnú definíciu.“

Ale predstavme si, že by som sa vedel vrátiť v čase na začiatok hádky:

(Eliezer sa vynára z ničoho v zvláštnej nádobe, ktorá vyzerá presne ako stroj času z pôvodného filmu Stroj času.)

Barry: „Do paroma! Cestovateľ v čase!“

Eliezer: „Som cestovateľ v čase! Počujte moje slová! Cestoval som do minulosti – približne pätnásť minút...“

Albert: „Pätnásť *minút*?“

Eliezer: „...aby som vám priniesol túto správu!“

(Nasleduje pauza, zmes zmätku a očakávania.)

Eliezer: „Myslíte si, že by sa ‚zvuk‘ mal definovať ako akustická vibrácia (tlakové vlny vo vzduchu) spojená so sluchovým vnemom (niekto ten zvuk počuje), alebo by sa ‚zvuk‘ mal definovať iba ako akustická vibrácia, prípadne iba ako sluchový vnem?“

Barry: „Ty si cestoval v čase, aby si sa nás opýtal *toto*?“

Eliezer: „Do mojich motívov vás nič nie je! Odpovedzte!“

Albert: „No... nevidím dôvod, prečo by na tom malo záležať. Môžeš si vybrať hocijakú definíciu, pokiaľ ju používaš konzistentne.“

Barry: „Hoď si mincou. Teda, hoď si mincou dvakrát.“

Eliezer: „Osobne by som povedal, že keď vyvstane tento problém, obe strany by mohli prepnúť na opisovanie udalosti pomocou jednoznačných zložiek na nižšej úrovni, ako sú akustické vibrácie alebo sluchové vnemy. Prípadne by si každá strana mohla vytvoriť nové slovo, napríklad ‚alberzle' a ‚bargulum', ktoré bude používať namiesto toho, čo predtým nazývali ‚zvuk'; a potom môžu obe strany používať tieto nové slová konzistentne. Potom žiadna strana nemusí ustupovať alebo stratiť tvár, ale stále môžu komunikovať. A samozrejme by ste si stále mali pamätať nejakú testovateľnú predpoveď, o čom tá debata naozaj je. Znie vám toto správne?"

Albert: „Asi..."

Barry: „Prečo sa o tomto bavíme?"

Eliezer: „Aby sa vaše priateľstvo zachovalo napriek skutočnostiam, o ktorých sa už nikdy nedozviete. Pretože budúcnosť sa už zmenila!"

(Eliezer a stroj zmiznú v obláčiku dymu.)

Barry: „Kde sme to vlastne boli?"

Albert: „Aha: Keď strom spadne v lese, a nikto ho nepočuje, vydá pritom zvuk?"

Barry: „Urobí pritom alberzle, ale nie bargulum. Čo je ďalšia otázka?"

Tento liek nezničí *všetky* debaty o kategorizácii. Ale zničí podstatnú časť z nich.

*→

164. Precíťte zmysel

Keď niekoho počujem povedať: „Aha, pozri, motýľ," vyslovené hlásky „motýľ" vstúpia do môjho ucha, rozochvejú môj ušný bubienok, prenesú sa na slimáka, pošteklia sluchové nervy, ktoré prenesú aktivačný impulz do sluchovej kôry, kde sa hlásky začnu spracovávať, spolu s rozoznávaním slov a rekonštrukciou syntaxe (toto vôbec nie je sériový proces), a mnoho ďalších komplikácií.

Ale na záver dňa, či skôr na záver sekundy, som nastavený pozrieť sa, kam môj kamarát ukazuje a vidieť zrakový vnem, ktorý rozoznám ako motýľa; a bol by som dosť prekvapený, keby som namiesto toho uvidel vlka.

Môj kamarát sa pozerá na motýľa, jeho hrdlo sa chveje a pery sa hýbu, tlakové vlny neviditeľne cestujú vzduchom, moje ucho počuje, moje nervy prenášajú, a môj mozog rekonštruuje, a hľa, viem, na čo sa môj kamarát pozerá. Nie je to úchvatné? Keby sme nevedeli o tlakových vlnách vo vzduchu, bol by to ohromný objav vo všetkých noviných: Ľudia ovládajú telepatiu! Ľudské mozgy dokážu navzájom prenášať myšlienky!

Áno, naozaj *máme* telepatiu; ale mágia nie je vzrušujúca, keď je iba *skutočná*, a keď to dokážu aj všetci vaši kamaráti.

Myslíte si, že telepatia je jednoduchá? Skúste postaviť počítač, ktorý bude telepatický voči vám. Telepatia, alebo „jazyk", alebo akokoľvek chcete nazvať našu čiastočnú schopnosť prenášať myšlienky, je zložitejšia než vyzerá.

Bolo by však pomerne nepohodlné chodiť okolo a myslieť si: „Teraz čiastočne preložím niektoré črty mojich myšlienok do lineárnej postupnosti hlások, ktorá vyvolá podobné myšlienky u môjho konverzačného partnera..."

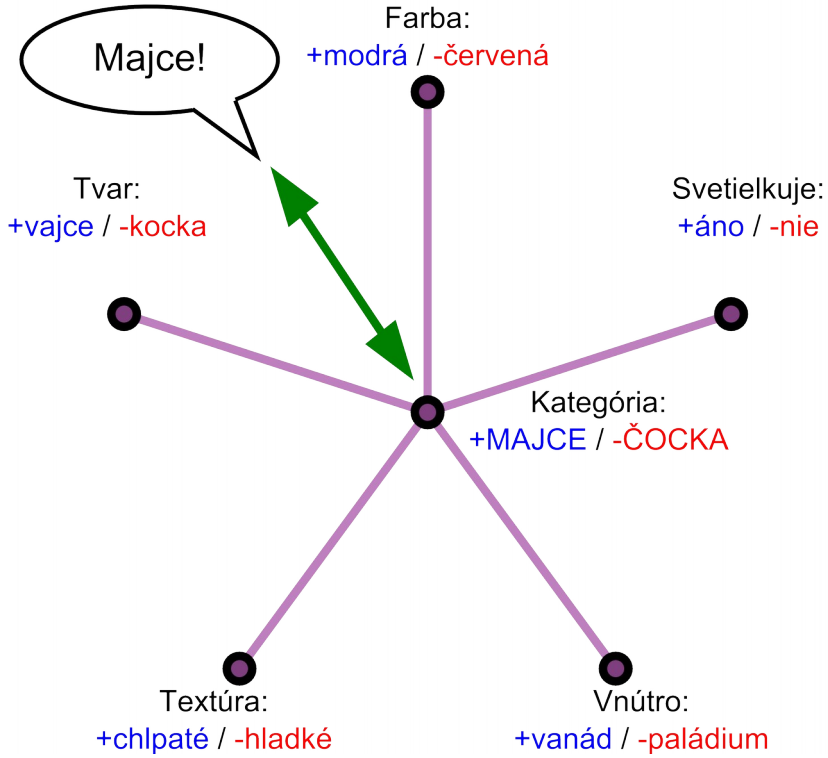
Preto mozog ukryje všetku zložitosť – alebo ju v prvom rade ani nikdy nereprezentuje – čo vedie ľudí k tomu, že si o slovách myslia pár zvláštnych vecí.

Ako som skôr poznamenal, keď na mňa vyskočí veľký žltý pásikavý predmet, pomyslím si: „Ej! Tiger!" a nie „Hm... predmety s vlastnosťami veľkosť, žltosť a pásikavosť mali v minulosti často aj

→ http://lesswrong.com/lw/np/disputing_definitions/

vlastnosti hladnosť a nebezpečnosť, a preto, hoci to nie je logicky nevyhnutné, *aúúúú* CHRUM CHRUM GLG.“

Podobne, keď niekto zakričí: „Ej! Tiger!“, prirodzený výber by neuprednostnil organizmus, ktorý by si pomyslel: „Hm... práve som počul slabiky ‚ti‘ a ‚ger‘, ktoré moji súkmeňovci spájajú so svojimi vnútornými analógiami môjho pojmu tigra, a ktoré s veľkou pravdepodobnosťou vyslovia, keď vidia predmet, ktorý zaradia do kategórie *aíííí* CHRUM CHRUM *pomoc odtrhlo mi to ruku* CHRUM GLG.“



Obrázok 164.1: Sieť 3

Keď zvážite toto dizajnové obmedzenie ľudskej poznávacej architektúry, nebudete chcieť mať *žiadne* ďalšie kroky medzi tým, keď vaša sluchová kôra rozozná slabiky „tiger“, a keď sa aktivuje pojem tigra.

Keď sa vrátime k podobenstvu o majciach a čockách, a centralizovanej sieti, ktorá kategorizuje rýchlo a lacno, môžete si predstaviť priame spojenie vedúce od jednotky, ktorá rozoznáva slabiky „majce“ k jednotke v strede majcovej siete. Centrálna jednotka, pojem majce, sa aktivuje takmer ihneď ako počujete seniorku triedičku Susan povedať „Majce!“

Alebo, za účelom rozprávania – ktoré by tiež nemalo trvať celé veky – akonáhle vidíte modrú vajcovitú vec, a centrálna jednotka majca vyšle signál, zakričíte na Susan: „Majce!“

A algoritmus to zvnútra cíti tak, že toto značenie a tento pojem sú takmer *totožné*; zmysel *pripadá ako* vnútorná vlastnosť samotného slova.

Sčítanejší to rozoznajú ako ďalší príklad „Klamu projekcie mysle“ E. T. Jaynesa. Zdá sa, akoby slovo *malo* význam, ako súčasť samotného slova; podobne ako červenosť sa zdá byť súčasťou jablka, alebo tajomnosť súčasťou tajomného javu.

Veru, vo väčšine prípadov mozog nebude vôbec rozlišovať medzi slovom a jeho významom – unúva sa tie dve veci oddeľovať iba kým sa učíte nový jazyk, možno. A ešte aj vtedy uvidíte, ako Susan ukazuje na modrú vajcovitú vec a hovorí „Majce!“ a pomyslíte si, *ktovie čo „majce“ znamená*, a nie *ktovie, akú myšlienkovú kategóriu si Susan spája so zvukovou značkou „majce“*.

Zvážte v tomto svetle časť Štandardnej Debaty o Definíciách, kde sa dve strany hádajú o tom, čo slovo „zvuk“ *naozaj* znamená – rovnako ako by sa mohli hádať, či je konkrétne jablko *naozaj* červené alebo zelené:

> Albert: „Mikrofón na mojom počítači dokáže nahrať zvuk, aj keď nablízko nie je nikto, kto by ho počul, uloží ho ako súbor, a to sa volá ‚zvukový súbor‘. A v tom súbore je uložený

vzor vibrácií vzduchu, nie vzor neurónových signálov v niekoho mozgu. ‚Zvuk' znamená vzor vibrácií."

Barry: „Vážne? Tak sa pozrime, či s tebou slovník súhlasí."

Albert intuitívne cíti, že slovo „zvuk" *má význam*, a že ten význam *sú* akustické vibrácie. Rovnako ako Albert cíti, že strom padajúci v lese *vydá zvuk* (a nie spôsobí udalosť, ktorá *zodpovedá kategórii zvuku*).

Barry podobne *cíti*, že:

```
zvuk.zmysel == sluchový vnem
les.zvuk == nepravda
```

Namiesto:

```
môjMozog.NájdiPojem(„zvuk") == pojem_SluchováSkúsenosť
pojem_SluchováSkúsenosť.zodpovedá(les) == nepravda
```

Čo je bližšie k tomu, čo sa *naozaj* odohráva; ale ľudia sa nevyvinuli, aby to vedeli, rovnako ako ľudia inštinktívne nevedia, že mozog sa skladá z neurónov.

Odporujúce si intuície Alberta a Barryho poskytujú palivo na pokračovanie hádky vo fáze hádania sa, čo slovo „zvuk" znamená - čo im *pripadá* ako hádanie sa o fakte ako je hocijaký iný fakt, napríklad hádanie sa o tom, či je obloha modrá alebo zelená.

Možno si ani nevšimnete, že niečo zablúdilo, dokiaľ sa nepokúsite vykonať racionalistický rituál stanovenia testovateľného pokusu, ktorého výsledok závisí na faktoch, o ktorých tak zapálene debatujete...

*→

165. Argumentovanie bežným používaním

Časť Štandardnej Debaty o Definíciách prebieha takto:

Albert: „Pozri, predpokladajme, že som nechal v lese mikrofón a nahral som vzor akustických vibrácií padajúceho stromu. Keby som to niekomu pustil, nazval by to ‚zvuk'! To je bežné použitie! Nevymýšľaj si svoje vlastné pochybné definície!"

Barry: „Po prvé, môžem definovať slovo tak, ako sa mi zachce, pokiaľ ho používam konzistentne. Po druhé, význam, ktorý som použil, je v slovníku. Po tretie, kto ti dal právo rozhodovať, čo je a čo nie je bežné použitie?"

Nie všetky debaty o definíciách sa dostanú tak ďaleko, aby si všimli pojem bežného použitia. Myslím si, že omnoho častejšie niekto vezme slovník, pretože verí, že slová majú významy, a že slovník verne zaznamenáva, čo je tento význam. Niektorí ľudia dokonca veria, že slovník *určuje* význam – že redaktori slovníkov sú Zákonodarcovia Jazyka. Možno preto, že im ešte na základnej škole ich učiteľská autorita povedala, že sa musia riadiť slovníkom, že je to povinné pravidlo a nie voliteľné?

Redaktori slovníkov čítajú, čo iní ľudia píšu, a zaznamenávajú, čo slová napohľad znamenajú; sú to historici. Oxfordský slovník anglického jazyka môže byť *súhrnný*, ale nie *autoritatívny*.

Ale iste existuje spoločenský príkaz používať slová všeobecne chápaným spôsobom? Nie je naša ľudská telepatia, naša vzácna sila jazyka, závislá na spoločnej koordinácii? Možno by sme mali dobrovoľne považovať redaktorov slovníkov za najvyšších rozhodcov – aj keď sa *oni* sami radšej považujú za historikov – aby sme udržali tichú spoluprácu, na ktorej závisí všetka reč.

Slovné spojenie „autoritatívny slovník" sa takmer nikdy nepoužíva správne; príkladom správneho použitia by bol *Autoritatívny slovník štandardov IEEE*. IEEE je skupina hlasujúcich členov, ktorí sa z profesionálnych dôvodov potrebujú presne zhodnúť na pojmoch a definíciách, preto je *Autoritatívny slovník štandardov IEEE* skutočnou dohodnutou legislatívou, ktorá vykonáva autoritu, ktorú pripisujeme IEEE.

→ http://lesswrong.com/lw/nq/feel_the_meaning/

V každodennom živote spoločný jazyk zvyčajne nevyrastá z vedomej dohody, ako je to v IEEE. Je to skôr otázkou nákazy ako sa slová vymýšľajú a šíria v kultúre. (Ako pred tridsiatimi rokmi povedal Richard Dawkins, sú to „mémy“ - ale už viete, čo tým myslím, a ak nie, môžete si to vygoogliť, a potom budete aj vy nakazení.)

Napriek tomu, ako ukazuje príklad IEEE, zhoda na jazyku môže byť spoločne vytvoreným verejným statkom. Ak si chceme vymieňať myšlienky pomocou jazyka, ľudskej telepatie, potom je v našom spoločnom záujme používať *rovnaké* slová na podobné pojmy – pokiaľ možno na pojmy čo najpodobnejšie v rámci rozlišovacích schopností našich mozgov – aj keď nemáme žiaden zrejmý spoločný záujem používať pre daný pojem nejaké *konkrétne* slovo.

Nemáme žiaden zrejmý spoločný záujem používať slovo „oto“ v zmysle zvuk, alebo „zvuk“ v zmysle oto; ale máme spoločný záujem používať *to isté* slovo, nech už je to ktorékoľvek z nich. (Bolo by lepšie, keby často používané slová boli kratšie, ale zatiaľ ešte nezachádzajme do teórie informácií.)

Ale aj keď máme spoločný záujem, nie je celkom *nevyhnutné*, aby sme vy a ja používali rovnaké označenia *vnútorne*; je to iba pohodlné. Ak viem, že pre vás „oto“ znamená zvuk – čiže, že si „oto“ spájate s pojmom veľmi podobným tomu, ktorý si ja spájam so „zvuk“ - potom môže povedať: „Papier pri krčení vydáva praskavý oto.“ Vyžaduje si to rozmýšľanie navyše, ale ak chcem, zvládnem to.

Podobne, ak poviete: „Čo je palica kolkovej gule dopadajúcej na podlahu?“ a ja viem, ktorý pojem si *vy* spájate so slabikami „palica“, dokážem zistiť, čo tým myslíte. Môže to vyžadovať trochu rozmýšľania a spomalí ma to, pretože ja si zvyčajne spájam „palica“ s iným pojmom. Ale zvládnem to.

Keď ľudia naozaj *chcú* navzájom komunikovať, je nás ťažké zastaviť! Keby sme boli zaseknutí na opustenom ostrove bez spoločného jazyka, vzali by sme palice a kreslili obrázky do piesku.

Albertovo odvolanie sa na argument bežným používaním predpokladá, že zhoda na jazyku je spoločne vytvorený verejný statok. Albert to však predpokladá iba za jediným účelom, rečnícky obviniť Barryho z porušovania dohody a ohrozovania verejného statku. Teraz sa už hádka o padajúcom strome celkom presunula z botaniky do politiky; a tak Barry reaguje spochybnením Albertovej autority definovať toto slovo.

Racionalista s aktívnou schopnosťou pritisnúť si otázku by si všimol, že konverzácia zablúdila veľmi ďaleko.

Aha, milý čitateľ, je to všetko naozaj potrebné? Albert vie, čo Barry myslí slovom „zvuk“. Barry vie, čo Albert myslí slovom „zvuk“. Aj Albert aj Barry poznajú slová ako „akustické vibrácie“ a „sluchové vnemy“, ktoré si už spájajú s rovnakým pojmami, a pomocou ktorých môžu opísať javy v lese bez nejednoznačnosti. Keby boli zaseknutí na opustenom ostrove a pokúšali sa komunikovať jeden s druhým, ich práca by už bola *hotová*.

Keď obe strany *vedia*, čo druhá strana *chce* povedať, a obe strany obviňujú druhú stranu z porušovania „bežného používania“, nech už sa bavia o čomkoľvek, je jasné, že *nepracujú na spôsobe, ako navzájom komunikovať.* Ale to je predsa celý úžitok, ktorý nám bežné používanie poskytuje.

Prečo by ste sa hádali o zmysel slova, dve strany doťahujúce sa tam a naspäť? Ak je to iba neprimerane zveličený konflikt o priestor slov, a nejde o nič iné, potom obe strany potrebujú iba vytvoriť dve nové slová a tie používať konzistentne.

Kategorizácia však často funguje ako skryté odvodzovanie a maskované otázky. Je ateizmus „náboženstvo“? Ak niekto argumentuje, že spôsoby uvažovania v ateizme sú na rovnakej úrovni ako spôsoby uvažovania používané v judaizme, alebo že ateizmus je na rovnakej úrovni ako islam čo sa týka spôsobovania násilia, potom majú jasný argumentačný záujem zlepiť to všetko dohromady do nejasnej sivej škvrny „viera“.

Alebo si vezmime boj za zmiešanie dokopy černochov a belochov ako „ľudí“. Toto by nebola vhodná doba na vytváranie dvoch slov – ide tu práve o tú myšlienku, že by ste ich z morálneho hľadiska nemali vnímať odlišne.

Ale ak ide o nejaký empirický výrok, *alebo* o nejaký morálny výrok, už sa nemôžete odvolávať na bežné použitie.

Ak je otázka o tom, ako zoskupiť podobné veci za účelom odvodzovania, budú empirické predpovede závisieť na otázke; čo znamená, že definície môžu byť *nesprávne*. Konflikt medzi predpoveďami nemožno rozhodnúť anketou o názoroch.

Ak chcete vedieť, či dať ateizmus do rovnakého zhluku ako náboženstvá veriace v nadprirodzené veci za účelom nejakého konkrétneho empirického odvodenia, na toto vám slovník neodpovie.

Ak chcete vedieť, či černosi sú ľudia, na toto vám slovník neodpovie.

Ak každý verí, že červené svetlo na oblohe je Mars, boh vojny, slovník bude definovať „Mars“ ako boha vojny. Ak každý verí, že oheň je uvoľňovanie flogistonu, slovník bude definovať „oheň“ ako uvoľňovanie flogistonu.

Používať slová je umenie; dokonca aj keď definície nie sú doslova pravdivé alebo nepravdivé, často sú múdrejšie alebo hlúpejšie. Slovníky sú iba dejinami minulého používania; ak ich beriete ako najvyšších rozhodcov významu, zväzuje vás to s múdrosťou minulosti, zakazuje vám to robiť veci lepšie.

Dávajte si však pozor na to (ak sa musíte odchýliť od múdrosti minulosti), aby ľudia dokázali zistiť, čo sa im pokúšate plávať.

*→

166. Prázdne nálepky

Vezmim si (áno, zase) aristotelovskú predstavu kategórií. Povedzme, že existuje nejaký predmet s vlastnosťami A, B, C, D a E, alebo prinajmenšom vyzerajúci trochu E.

Fred: „Myslíš tým, že tá vec je modrá, guľatá, chlpatá a...“

Ja: „Podľa aristotelovskej logiky by nemal byť rozdiel v tom, aké vlastnosti to sú, alebo ako ich volám. Práve preto používam písmená.“

Ďalej vymyslím aristotelovskú kategóriu „zawa“, ktorá popisuje tie predmety, všetky také predmety a iba také predmety, ktoré majú vlastnosti A, C a D.

Ja: „Predmet 1 je zawa, B a E.“

Fred: „Ale je aj modrý... teda A... však?“

Ja: „To vyplýva z toho, keď poviem, že je zawa.“

Fred: „Aj tak by som bol radšej, keby si to povedal priamo.“

Ja: „Dobre. Predmet 1 je A, B, zawa a E.“

Potom pridám ďalšie slovo, „yokie“, ktoré opisuje všetky a iba tie predmety, ktoré sú B aj E; a slovo „xippo“, ktoré popisuje všetky a iba tie predmety, ktoré sú E ale nie D.

Ja: „Objekt 1 je zawa a yokie, ale nie xippo.“

Fred: „Počkaj, je priesvitný? Chcem povedať, je E?“

Ja: „Áno. To je jediná možnosť pri daných informáciách.“

Fred: „Radšej by som bol, keby si to vymenoval.“

Ja: „Dobre: Predmet 1 je A, zawa, B, yokie, C, D, E, ale nie xippo.“

Fred: „Úžasné! Toto všetko dokážeš zistiť púhym pohľadom?“

Pôsobivé, nie? Vymyslime ešte viac nových slov: „olo“ je A, C a yokie; „mun“ je A, C a xippo; a „merlacdonský“ je olo a mun.

Zbytočne mätúce? Tiež si myslím. Poďme nahradiť nálepky definíciami:

„Zawa, B a E“ bude [A, C, D], B, E

„Olo a A“ bude [A, C, [B, E]], A

→ http://lesswrong.com/lw/nr/the_argument_from_common_usage/

„Merlacdonský“ bude [A, C, [B, E]], [A, C, [E, ~D]]

A dôležité je zapamätať si o aristotelovskej predstave kategórií, že [A, C, D] je *všetka* informácia o „zawa“. Neznamená to len, že môžem túto nálepku zmeniť, ale že by som sa rovnako dobre zaobišiel aj celkom bez nálepiek – pravidlá pre aristotelovské triedy fungujú čisto na štruktúrach ako je [A, C, D]. Nazvať niektorú z týchto štruktúr „zawa“, alebo vôbec k nej priradiť nejakú nálepku, je iba ľudská skratka (alebo komplikácia), ktorá pre aristotelovské pravidlá neznamená žiaden rozdiel.

Povedzme, že „človek“ definujeme ako smrteľný dvojnožec bez peria. Potom má klasický sylogizmus podobu:

Všetci [smrteľný, dvojnožec, ~perie] sú smrteľní.

Sokrates je [smrteľný, dvojnožec, ~perie].

Preto, Sokrates je smrteľný.

Táto zručnosť usudzovania odrazu vyzerá omnoho menej pôsobivo, nie?

Ilúzia odvodenia tu vychádza z nálepiek, ktoré ukrývajú predpoklady a predstierajú, že v závere je niečo nové. Nahradenie nálepiek definíciami odhaľuje túto ilúziu, zviditeľňuje empirickú neužitočnosť tautológie. Nikdy nemôžete povedať, že Sokrates je [smrteľný, dvojnožec, ~perie], dokiaľ ste nepozorovali, že je smrteľný.

Existuje predstava, ktorú, ako ste si už možno všimli, neznášam, že „môžete definovať slovo ako sa vám zachce“. Táto predstava vychádza z aristotelovskej predstavy kategórií; lebo ak sa riadite aristotelovskými pravidlami *presne* a *bezchybne* – čo ľudia nikdy nerobia; Aristoteles veľmi dobre vedel, že Sokrates je človek, hoci na to podľa svojich vlastných pravidiel nemal nárok – ale *ak* by sa nejaká fiktívna bytosť pokúšala riadiť tými pravidlami presne, nikdy by nedošla k rozporu. Nikdy by totiž nedošla k ničomu: nemohla by povedať, že Sokrates je [smrteľný, dvojnožec, ~perie], dokiaľ by nepozorovala, že je smrteľný.

Ale nejde tu až tak o to, že nálepky v aristotelovskom systéme sú *svojvoľné*, ako že aristotelovský systém funguje dobre aj *bez akýchkoľvek nálepiek* – chŕli presne ten istý prúd tautológií, akurát vyzerajú menej pôsobivo. Nálepky sú tam iba na vytvorenie *ilúzie* odvodzovania.

Ak teda vôbec chcete mať nejaké aristotelovské porekadlo, to porekadlo by nemalo byť „môžem definovať slovo, ako sa mi zachce“, dokonca ani „definovanie slova nemá žiadne dôsledky“, ale skôr: „definície nepotrebujú slová“.

*→

167. Hrajte so slovami tabu

V hre Tabu (od firmy Hasbro) je cieľom hráča, aby jeho partner uhádol slovo napísané na karte, bez použitia tohto slova alebo ďalších piatich slov uvedených na karte. Napríklad môžete mať za úlohu, primäť vášho partnera povedať „baseball“ bez použitia slov „šport“, „pálka“, „odpal“, „nahadzovač“, „méta“ a samozrejme „baseball“.

Keď vidím takúto úlohu, hneď si pomyslím: „Umelý skupinový konflikt, pri ktorom používate dlhý drevený valec, ktorým udriete do hodeného guľovitého telesa, a potom bežíte cez štyri bezpečné miesta.“ Možno to nie je tá najefektívnejšia stratégia na sprostredkovanie slova „baseball“ pri daných pravidlách – to by mohlo byť: „čo hrávajú Yankeeovia“ - ale všeobecnú zručnosť *vymazania slova z mysle* cvičím už roky, hoci z iného dôvodu.

Včera sme videli ako nahradenie pojmu jeho definíciou môže odhaliť empirickú neproduktívnosť klasického aristotelovského sylogizmu. Všetci ľudia sú smrteľní (a zároveň údajne aj dvojnožci bez peria); Sokrates je človek; preto Sokrates je smrteľný. Keď nahradíme slovo „človek“ jeho údajnou definíciou, odhalí sa za tým nasledujúce uvažovanie:

Všetci [smrteľný, dvojnožec, ~perie] sú smrteľní;

→ http://lesswrong.com/lw/ns/empty_labels/

Sokrates je [smrteľný, dvojnožec, ~perie];

Preto, Sokrates je smrteľný.

Princíp nahrádzania slov definíciami sa však dá použiť omnoho širšie:

Albert: „Strom padajúci v opustenom lese vydá zvuk.“

Barry: „Strom padajúci v opustenom lese nevydá zvuk.“

Je jasné, že keď jeden hovorí „zvuk“ a druhý hovorí „nie zvuk“, musíme mať rozpor, však? Ale predpokladajme, že by obaja pred hovorením odreferencovali svoje ukazovatele:

Albert: „Strom padajúci v opustenom lese zodpovedá [test členstva: táto udalosť vytvára akustické vibrácie].“

Barry: „Strom padajúci v opustenom lese nezodpovedá [test členstva: táto udalosť vytvára sluchové vnemy].“

Teraz tu už nie je domnelá zrážka – jediné, čo bolo treba urobiť, je zakázať si používať slovo *zvuk*. Keby vznikla hádka o „akustické vibrácie“, môžeme si opäť zahrať Tabu a povedať „tlakové vlny v hmotnom prostredí“; keby bolo treba, môžeme si zahrať Tabu aj pre slovo „vlna“ a nahradiť ho vlnovou rovnicou. (Keď zahráme Tabu na „sluchové vnemy“, dostaneme „tá forma zmyslového vnímania v ľudskom mozgu, ktorá dostáva ako vstup lineárnu časovú postupnosť zmesi frekvencií...“)

Predstavme si teraz naopak, že by Albert a Barry mali hádku:

Albert: „Sokrates zodpovedá [test členstva: táto osoba zomrie, keď vypije bolehlav].“

Albert: „Sokrates zodpovedá [test členstva: táto osoba nezomrie, keď vypije bolehlav].“

Teraz majú Albert a Barry podstatný konflikt v očakávaniach; rozdiel v tom, čo očakávajú, že uvidia, keď Sokrates vypije bolehlav. Nemuseli by si to však uvedomiť, keby používali to isté slovo „človek“ na rôzne pojmy.

Dostanete veľmi odlišný pohľad na to, v čom ľudia súhlasia alebo nesúhlasia, podľa toho, či sa pozeráte z pohľadu nálepky (Albert hovorí „zvuk“ a Barry hovorí „nie zvuk“, takže nesúhlasia) alebo či sa pozeráte z pohľadu testu (Albertov test členstva sú akustické vibrácie, Barryho je sluchový vnem).

Dajte dokopy skupinu *takzvaných* futuristov a opýtajte sa ich, či si myslia, že o tridsať rokov budeme mať umelú inteligenciu, a tipujem, že aspoň polovica z nich povie áno. Ak to takto necháte, potrasú si rukami a navzájom si zablahoželajú ku konsenzu. Ale vyhláste pojem „umelá inteligencia“ za tabu, a požiadajte ich, aby opísali, *čo* očakávajú, že uvidia, bez použitia slov ako „počítač“ a „myslieť“, a možno nájdete veľké konflikty očakávaní skryté pod týmto beztvarým štandardným slovom. A pozrite si Shaneovu Leggovu zbierku 71 definícií „inteligencie“.

Ilúziu jednoty medzi náboženstvami dokážete zrušiť tým, že vyhlásite pojem „Boh“ za tabu, a požiadate ich, aby povedali, čo je to, v čo veria; alebo vyhláste za tabu slovo „viera“ a opýtajte sa ich, prečo si to myslia. Väčšina z nich však vôbec nebude schopná odpovedať, pretože je to v prvom rade najmä vyznávanie, a nedokážete kognitívne zaostriť na zvukovú nahrávku.

Keď zistíte, že máte filozofické problémy, *prvá obranná línia nie je definovať svoje problematické pojmy, ale pozrieť, či dokážete rozmýšľať aj bez používania týchto pojmov.* A ich skratiek a synoným. Dávajte si pozor, aby ste namiesto toho nevymysleli nové slovo. Opíšte vonkajšie pozorovania a vnútorné mechanizmy; nepoužívajte jeden ovládač, nech už je tým ovládačom čokoľvek.

Albert hovorí, že ľudia majú „slobodnú voľu“. Barry hovorí, že ľudia nemajú „slobodnú vôľu“. Nuž, toto isto vyvolá domnelý konflikt. Väčšina filozofov by Albertovi a Barrymu poradila, aby skúsili presne definovať, čo myslia slovami „slobodná vôľa“, a na túto tému iste dokážu diskutovať veľmi dlho. Ja by som Albertovi a Barrymu poradil, aby opísali, čo je to, čo si myslia, že ľudia majú alebo nemajú, celkom bez použitia slov „slobodná vôľa“. (Ak si to chcete vyskúšať doma, odporúčam vyhnúť sa aj slovám „výber“, „konať“, „rozhodnúť sa“, „určené“, „zodpovedné“ a akékoľvek ich synonymá.)

Toto je jeden z neštandardných nástrojov v mojej zbierke, a podľa môjho skromného úsudku funguje *omnoho omnoho* lepšie než ten štandardný. Takisto si jeho použitie vyžaduje viac úsilia; dostanete to, za čo si zaplatíte.

⁂→

168. Nahraďte symbol podstatou

Čo treba na to, aby ste – ako v príklade v predchádzajúcej kapitole – videli „baseballový zápas“ ako „umelý skupinový konflikt, pri ktorom používate dlhý drevený valec, ktorým udriete do hodeného guľovitého telesa, a potom bežíte cez štyri bezpečné miesta“? Čo treba na hranie racionalistickej verzie Tabu, ktorej cieľom nie je nájsť synonymum, ktoré nie je na karte, ale nájsť spôsob opisu bez použitia štandardného uchopenia pojmu?

Musíte vizualizovať. Musíte svojou mysľou vidieť podrobnosti, akoby ste sa pozerali po prvýkrát. Musíte vykonať pôvodné videnie.

Je toto „pálka“? Nie, je to dlhá, zaguľatená, zužujúca sa drevená tyč, zúžená na jednom konci tak, že ju človek môže chytiť a mávnuť ňou.

Je toto „lopta“? Nie, je to kožou pokryté guľovité teleso so súmerným vzorom šitia, pevné, ale nie také pevné ako kov, ktoré môže niekto chytiť a hodiť, alebo udrieť tou drevenou tyčou, alebo chytiť.

Sú toto „méty“? Nie, sú to dané miesta na ihrisku, kam sa hráči snažia čo najrýchlejšie utekať, pretože im to v rámci umelých pravidiel hry poskytuje bezpečie.

Hlavou prekážkou vykonania pôvodného videnia je, že vaša myseľ už má pekný úhľadný súhrn, pekné malé jednoducho použiteľné držadlo na pojem. Ako slovo „baseball“ alebo „pálka“ alebo „méta“. Vyžaduje to úsilie zastaviť svoju myseľ, aby nekĺzala po známej ceste, tej ľahkej ceste, ceste najmenšieho odporu, kde malé beztvaré slovo vtrhne a zničí podrobnosti, ktoré sa pokúšate vidieť. Samotné slovo môže mať deštruktívnu silu klišé; samotné slovo môže niesť jed uloženej myšlienky.

Hrať hru Tabu – dokázať opísať bez použitia štandardného ukazovateľa/nálepky/držadla – je jedna zo *základných* racionalistických schopností. Nachádza sa na rovnakej prvotnej úrovni ako zvyk ustavične sa pýtať „Prečo?“ alebo „Čo na základe tohto názoru očakávam?“

Toto umenie úzko súvisí s:

- Pragmatizmom, pretože videnie týmto spôsobom vás omnoho lepšie spája s očakávanou skúsenosťou, namiesto výrokového názoru;
- Redukcionizmom, pretože videnie týmto spôsobom vás často núti zísť na nižšiu úroveň usporiadania, pozrieť sa na časti namiesto letmého pohľadu na celok;
- Pritisnutím si otázky, pretože slová často odpútavajú pozornosť od otázky, ktorú naozaj chcete položiť;
- Vyhýbaním sa uloženým myšlienkam, ktoré vtrhnú pri používaní bežných slov, takže ich môžete zablokovať zakázaním bežných slov;
- Spisovateľským pravidlom „Ukáž, nehovor!“, ktoré má medzi racionalistami moc;
- A nespúšťaním svojho pôvodného cieľa zo zreteľa.

Ako vám zakazovanie slova môže pomôcť udržať si svoj cieľ?

Z kapitoly Stratené ciele:

> Vo chvíli, keď čítate tento text, nejaký mladý muž alebo žena sedí na univerzite v lavici, usilovne študuje materiál, ktorý vôbec nemá v úmysle niekedy použiť, a ani sa oň nezaujíma kvôli nemu samotnému. Chce dobre platenú prácu, dobre platená práca vyžaduje kus papiera, ten kus papiera vyžaduje ukončenie magisterského štúdia, magisterské štúdium vyžaduje ukončenie bakalárskeho štúdia, a univerzita poskytujúca bakalárske štúdium vyžaduje, aby študenti absolvovali kurz vzorov pletenia v 12. storočí. Preto usilovne študuje, so zámerom

→ http://lesswrong.com/lw/nu/taboo_your_words/

zabudnúť to všetko v okamihu, keď dokončí záverečnú skúšku, ale aj tak vážne pracuje, pretože *chce* ten kus papiera.

Načo chodíte do „školy“? Aby ste dostali „vzdelanie“ ukončené „diplomom“. Vymažte zakázané slová a ich zrejmé synonymá, predstavte si skutočné podrobnosti, a máte omnoho väčšiu šancu si všimnúť, že „škola“ dnes znamená sedieť vedľa unudených teenagerov počúvajúcich veci, ktoré dávno viete, že „diplom“ je kus papiera, na ktorom je niečo napísané, a že „vzdelanie“ je zabudnutie tohto učiva akonáhle vás z neho doskúšajú.

Presakujúce zovšeobecnenia sa často prejavujú pomocou kategorizácií: Ľudia, ktorí sa v triede naozaj učia, sú zaradení do kategórie „vzdelávajú sa“, preto „vzdelávanie sa“ musí byť dobré; ale potom každý, kto naozaj príde do školy, tiež zapadne pod hlavičku „vzdeláva sa“ bez ohľadu na to, či sa učí alebo nie.

Študenti, ktorí rozumejú matematike, budú dobrí v testoch, ale ak vyžadujete, aby školy produkovali dobré skóre v testoch, budä tráviť všetko svoj čas učením na testy. *Myšlienková kategória*, ktorá nedokonale vystihuje váš cieľ, dokáže spôsobiť rovnakú motivačnú chybu *vo vnútri*. Chcete sa učiť, takže potrebujete „vzdelanie“; a dokiaľ dostávate hocičo, čo spadá do kategórie „vzdelanie“, nemusíte si všimnúť, či sa učíte alebo nie. Alebo si všimnete, ale neuvedomíte si, že ste stratili zo zreteľa svoj pôvodný cieľ, pretože teraz dostávate „vzdelanie“ a tak ste si v myšlienkach opísali svoj cieľ.

Kategorizovať znamená odhadzovať informácie. Ak vám povedia, že padajúci strom vydá „zvuk“, neviete, aký zvuk to naozaj je; nepočuli ste naozaj ten strom padať. Ak na minci padne „hlava“, nepoznáte jej radiálnu orientáciu. Modrá vajcovitá vec môže byť „majce“, ale čo ak sa presný vajcovitý tvar mení, alebo presný odtieň modrej? Chcete používať kategórie, aby ste odhodili nepodstatné informácie, aby ste preosiali zlato z piesku, ale štandardné kategórie často odhadzujú aj podstatné informácie. A keď skončíte v takomto type mentálneho problému, prvé a najsamozrejmejšie riešenie je hrať Tabu.

Napríklad aj samotné „hrať Tabu“ je presakujúce zovšeobecnenie. Verzia firmy Hasbro nie je racionalistická verzia; oni iba na karte vymenujú päť zakázaných slov, a to zdaleka nestačí pokryť vylúčenie myslenia pomocou starých známych slov. To, čo robia racionalisti, by sa tiež rátalo ako hranie Tabu – spadalo by to pod ich predstavu „hrať Tabu“ - ale nie všetko, čo sa ráta ako hranie Tabu, tlačí k pôvodnému videniu. Ak si iba myslíte „hraním Tabu sa dotlačím k pôvodnému mysleniu“, začnete si myslieť, že všetko, čo sa počíta ako hranie Tabu, sa musí počítať ako pôvodné videnie.

Racionalistická verzia nie je hra, čo znamená, že nemôžete vyhrať tým, že budete chytrý a ohnete pravidlá. Musíte hrať Tabu s dobrovoľným hendikepom: Zakážte si používať synonymá, ktoré nie sú na karte. Musíte si aj zakázať vymyslieť jednoduché nové slovo alebo slovné spojenie, ktoré by fungovalo ako mentálne držadlo ekvivalentné tomu starému. Pokúšate sa zaostriť na svoju mapu, nie premenovať mestá; dereferencovať ukazovateľ, nie alokovať nový ukazovateľ; vidieť udalosti, ako sa dejú, nie prepísať klišé do odlišných slov.

Vizualizovaním si problému vo väčších podrobnostiach môžete uvidieť stratený cieľ: Čo presne robíte, keď „hráte Tabu“? Akému cieľu slúži každá jedna časť?

Ak vidíte svoje činnosti a situáciu pôvodne, dokážete pôvodne vidieť aj svoje ciele. Ak dokážete vidieť čerstvými očami, akoby po prvýkrát, zbadáte sa, že robíte veci, ktoré by vám na um neprišlo robiť, keby to neboli zvyky.

Cieľ sa stráca vždy, keď podstatu (učenie, vedomosti, zdravie) nahradíme symbolom (titul, body v teste, lekárska starostlivosť). Aby ste uzdravili stratený cieľ, alebo stratovú kategorizáciu, musíte urobiť opak:

Nahraďte symbol podstatou; nahraďte označovateľ označeným; nahraďte vlastnosť testom členstva; nahraďte slovo významom; nahraďte nálepku pojmom; nahraďte súhrn podrobnosťami; nahraďte zástupnú otázku skutočnou otázkou; dereferencujte ukazovateľ; zíďte na nižšiu úroveň usporiadania; myšlienkovo simulujte proces namiesto pomenovania; zaostrite na svojej mape.

„Jednoduchá pravda“ vznikla ako cvičenie v tejto disciplíne pri opise „pravdy“ na nižšej úrovni usporiadania, bez použitia pojmov ako „presný“, „správny“, „reprezentovať“, „odrážať“, „sémantický“, „myslieť si“, „vedomosť“, „mapa“ alebo „skutočný“. (A pamätajte, že *skutočným* cieľom nie je hrať Tabu – slovo „pravda“ sa v texte vyskytuje, ale *nie* na definovanie pravdy. V hre firmy Hasbro by vás za to vypískali, ale my *naozaj* nehráme túto hru. Opýtajte sa sami seba, či ten dokument splnil svoj cieľ, nie či dodržiaval pravidlá.)

Samotné Bayesovo pravidlo opisuje „indíciu“ čisto matematicky, bez použitia slov ako „vyplýva“, „znamená“, „podporuje“, „dokazuje“ alebo „zdôvodňuje“. Pokúste sa *definovať* takéto filozofické pojmy, a akurát sa budete točiť v kruhu.

A potom si treba zahrať Tabu s tým najdôležitejším slovom. Často som vás varoval, aby ste si dali pozor, nech ho príliš nepoužívate, alebo dokonca, aby ste sa v istých prípadoch tomuto pojmu vyhli. Teraz viete skutočný dôvod, prečo. Nie je to zlá téma na rozmýšľanie. Ale vaše skutočné pochopenie sa meria vašou schopnosťou opísať, čo robíte a prečo, *bez* použitia tohto slova alebo niektorého z jeho synoným.

*→

169. Chyby kompresie

Ako sa hovorí: „Mapa nie je územie.“ Jediná 100% presná mapa Kalifornie v životnej veľkosti, na atóm presná, je Kalifornia samotná. Ale Kalifornia má dôležité pravidelnosti, napríklad na opísanie tvaru jej diaľníc stačí oveľa menej informácií – a tiež oveľa menej menej *fyzického materiálu* – než by si vyžadovalo opísanie každého atómu vnútri hraníc štátu. Preto sa hovorí *aj*: „Mapa nie je územie, ale územie si nemôžeš poskladať a vložiť do priehradky na rukavice.“

Papierová mapa Kalifornie v škále 10 kilometrov na 1 centimeter (milión k jednej) nemá miesto na to, aby ukázala rôzne polohy dvoch padnutých listov ležiacich centimeter vedľa seba na chodníku. Aj keby sa mapa tieto listy pokúsila ukázať, listy by sa na mape objavili na tom istom mieste; presnejšie, mapa by potrebovala rozlíšenie 10 nanometrov, čo je menej než zvládne väčšina tlačiarní, nehovoriac o ľudských očiach.

Skutočnosť je veľmi veľká – už len tá časť, ktorú vidíme, má krížom miliardy svetelných rokov. Ale vaša mapa skutočnosti je napísaná na pár kilách neurónov poskladaných, aby sa zmestili do vašej lebky. Nechcem urážať, ale vaša lebka je maličká, keď to takto porovnáme.

Je teda nevyhnutné, že niektoré veci, ktoré sa v skutočnosti odlišujú, budú na vašej mape stlačené do toho istého bodu.

Ale zvnútra vám toto nepripadá tak, že poviete: „Aha, pozri, stláčam dve veci do jedného bodu na mojej mape.“ Zvnútra to *pripadá* tak, že je iba *jedna* vec, a vy ju vidíte.

Dostatočne malé dieťa alebo dostatočne staroveký grécky filozof nebude vedieť, že existujú také veci ako „akustické vibrácie“ a „sluchové vnemy“. Existuje iba jedna vec, ktorá sa stane, keď strom padne; jediná udalosť nazvaná „zvuk“.

Uvedomiť si, že existujú *dve* rôzne udalosti za *jedným* bodom na vašej mape, je v podstate *vedecká* úloha – veľká, náročná vedecká úloha.

Niekedy sú chyby kompresie výsledkom pomýlenia si dvoch známych vecí pod rovnakou nálepkou – poznáte akustické vibrácie, a viete o sluchovom spracovaní v mozgu, ale obe nazývate „zvuk“ a tak sa popletiete. Ale nebezpečnejšia chyba kompresie vzniká, keď *ani len netušíte*, že vôbec *existujú* dve rôzne veci. V triediacom systéme je iba jeden mentálny spis označený „zvuk“ a všetko ohľadom „zvuku“ ide do tohto spisu. Nie sú tam dva spisy s rovnakou nálepkou; iba jeden spis. Mapa je štandardne stlačená; prečo by mozog vytváral dve mentálne vedrá, keď stačí jedno?

Alebo pomyslite na detektívku, v ktorej je kritickým detektívovým zistením, že jeden z podozrivých má rovnaké dvojča. Pri bežnej detektívovej práci je jeho úlohou iba pozorovať, že Carol je oblečená v

→ http://lesswrong.com/lw/nv/replace_the_symbol_with_the_substance/

červenom, že má čierne vlasy, a kožené sandále – ale toto všetko sú *fakty o* Carol. Je pomerne jednoduché pochybovať o jednotlivom fakte, ako NosíČervenú(Carol) alebo ČierneVlasy(Carol). Možno je ČierneVlasy(Carol) nepravda. Možno si Carol farbí vlasy. Možno platí HnedéVlasy(Carol). Treba však rafinovanejšieho detektíva, aby sa zamyslel, či Carol v NosíČervenú(Carol) a ČierneVlasy(Carol) – v spise Carol, do ktorého padajú jeho pozorovania – netreba rozdeliť na *dva* spisy. Možno existujú dve Carol, takže Carol, ktorá nosí červenú, nie je tá istá žena ako Carol, ktorá má čierne vlasy.

Tu je samotný čin *vytvorenia* dvoch odlišných vedier zábleskom geniálneho osvietenia. Je jednoduchšie pochybovať o svojich faktoch než o svojej ontológii.

Mapa skutočnosti uložená v ľudskom mozgu, na rozdiel od papierovej mapy Kalifornie, sa môže dynamicky rozširovať, keď do nej zapisujeme podrobnejšie popisy. Ale zvnútra to ani tak nepripadá ako zaostrenie na mapu, ale skôr ako rozštiepenie nedeliteľného atómu – vziať *jednu vec* (pripadá to ako jedna vec) a rozdeliť ju na dve alebo viac vecí.

Toto sa často prejavuje vytvorením nových slov, ako „akustické vibrácie“ a „sluchovné vnemy“ namiesto iba „zvuk“. Niečo vo vytvorení nového mena asi vyhradzuje nové vedro. Detektív bude mať sklon začať volať jednu z jeho podozrivých „Carol 2“ alebo „tá druhá Carol“, takmer hneď ako si uvedomí, že sú dve.

Ale rozšíriť mapu nie je vždy také jednoduché ako vytvoriť nové názvy miest. Je zábleskom vedeckého osvietenia uvedomiť si, že niečo také ako akustické vibrácie alebo sluchové vnemy vôbec *existuje*.

Samozrejmý moderný príklad by boli slová ako „inteligencia“ alebo „vedomie“. Každú chvíľu vidíme tlačovú správu tvrdiacu, že nejaký výskum „vysvetlil vedomie“, pretože tím neurológov skúmal elektrický rytmus 40 Hz, ktorý by mohol nejako súvisieť s krížovou modalitou spájania zmyslových informácií, alebo pretože skúmali retikulárny aktivačný systém, ktorý udržiava ľudí bdelých. Toto je extrémny príklad, zvyčajné chyby sú jemnejšie, ale sú rovnakého druhu. Časť „vedomia“, ktorú ľudia považujú za najzaujímavejšiu, je reflektivita, sebauvedomovanie, uvedomenie si, že osoba, ktorú vidím v zrkadle, je „ja“; toto je ťažký problém subjektívnej skúsenosti, ako ho odlíšil Chalmers. Ako „vedomie“ označujeme aj stav, keď bdieme a nespíme v našom dennom cykle. Ale to sú všetko rôzne pojmy vyskytujúce sa pod rovnakým názvom, a javy za nimi sú rôzne vedecké hlavolamy. Dokážete vysvetliť bdenie bez vysvetlenia reflektivity alebo subjektivity.

Chyby kompresie sú aj za technikou chytáku vo filozofii – hádate sa o „vedomí“ podľa jednej definície (napríklad schopnosť myslieť o myslení) a potom použijete záver na „vedomie“ podľa inej definície (napríklad subjektivita). Samozrejme, je možné, že tie dve sú to isté, ale ak áno, skutočné *pochopenie* tohto faktu by si vyžadovalo *najprv* pojmové rozdelenie a *potom* geniálny záblesk opätovného spojenia.

Rozširovanie vašej mapy je (opakujem) *vedecká* úloha: časť umenia vedy, schopnosť pýtať sa na svet. (A samozrejme nemôžete vedeckú úlohu vyriešiť odvolávaním sa na slovníky, ani zvládnuť zložitú schopnosť pýtania sa slovami „môžem si definovať slovo, ako sa mi zachce“.) Kde vidíte jedinú mätúcu vec, s premenlivými a protirečivými vlastnosťami, je dobrý odhad, že vaša mapa toho tlačí príliš veľa do jedného bodu – že ho potrebujete roztiahnuť a vyhradiť si nejaké nové vedierka. To nie je ako *definovať* jednu vec, ktorú vidíte, ale často to *vyplýva* zo zisťovania, ako hovoriť o danej veci bez používania jedného myšlienkového držadla.

Zručnosť rozťahovania mapy je teda spojená s racionalistickou verziou hry Tabu, a s múdrym používaním slov; pretože slová často predstavujú body na našej mape, nálepky, pod ktorými si zatrieďujeme svoje výroky, a vedrá, do ktorých vhadzujeme svoje informácie. Vyhnúť sa jednému slovu alebo vyhradiť si nové slová je často súčasťou zručnosti rozširovania mapy.

*→

→ http://lesswrong.com/lw/nw/fallacies_of_compression/

170. Kategórie majú následky

Medzi mnohými genetickými variáciami a mutáciami, ktoré nosíte vo svojom genóme, existuje zopár alel, ktoré asi poznáte – vrátane tých, ktoré určujú vašu krvnú skupinu: prítomnosť alebo neprítomnosť antigénov A, B a +. Ak dostanete krvnú transfúziu obsahujúci antigén, ktorý nemáte, vyvolá alergickú reakciu. Karl Landsteiner objavil tento fakt, aj ako testovať kompatibilné krvné skupiny, a umožnil tak transfúziu krvi bez zabitia pacienta (Nobelova cena za medicínu 1930). Ďalej, ak matka s krvnou skupinou A (napríklad) vynosí dieťa s krvnou skupinou A+, matka môže získať alergickú reakciu na antigén +; keby mala ďalšie dieťa s krvnou skupinou A+, toto dieťa by bolo ohrozené, pokiaľ matka nebude počas tehotenstva brať látky na potlačenie alergie. Preto sa ľudia pýtajú na svoje krvné skupiny pred sobášom.

Aha, a *ešte*: ľudia s krvnou skupinou A sú vážni a tvoriví, zatiaľ čo ľudia s krvou skupinou B sú divokí a veselí. Ľudia so skupinou 0 sú príjemní a spoločenskí, kým ľudia so skupinou AB sú pohodoví a vyrovnaní. (Mohli by ste si myslieť, že 0 je neprítomnosť A a B, kým AB bude skrátka A plus B, ale nie...) Toto všetko je podľa japonskej teórie osobnosti podľa krvnej skupiny. Zdá sa, že krvná skupina hrá v Japonsku rovnakú rolu ako znamenia zvieratníka na Západe, vrátane horoskopov podľa krvnej skupiny v dennej tlači.

Táto móda je mimoriadne čudná, pretože krvné skupiny *nikdy neboli* tajomné, ani v Japonsku, ani nikde ine. O samotnej *existencii* krvných skupín vieme iba vďaka Karlovi Landsteinerovi. Žiaden mystický šaman, žiaden ctihodný čarodejník nikdy nepovedal jediné slovo o krvných skupinách; neexistujú žiadne prastaré zaprášené zvitky, ktoré by zahalili túto chybu do aury staroveku. Keby lekári zajtra oznámili, že to celé bol obrovský podvod, my laici by sme nemali jediný kúsok indície dostupnej našim holým zmyslom, aby sme im protirečili.

Nikdy nebola žiadna vojna medzi krvnými skupinami. Nikdy nebol ani len politický konflikt medzi krvnými skupinami. Tieto stereotypy museli vzniknúť *čisto* zo *samotnej existencie* nálepiek.

Teraz niekto určite poznamená, že toto je príbeh o kategorizovaní ľudí. Stane sa tá istá vec, ak kategorizujete rastliny, alebo kamene, alebo kancelársky nábytok? Nespomínam si, že by som čítal o takomto pokuse, ale to samozrejme neznamená, že taký nebol. (Očakával by som, že najťažšou časťou robenia takého experimentu by bolo nájsť protokol, ktorý by nezavádzal pokusné osoby do predstavy že, keďže vám dali túto nálepku, musí byť nejako dôležitá.) Takže hoci nechcem aktualizovať na základe imaginárnej indície, predpovedal by som, že pokus bude mať pozitívny výsledok: očakával by som, že zistia, že samotné nálepkovanie má moc nad všetkými vecami, prinajmenšom v ľudskej predstavivosti.

Môžete to vidieť v pojmoch zhlukov podobnosti: akonáhle nakreslíte hranicu okolo nejakej skupiny, myseľ začne skúšať zbierať podobnosti v skupine. A nanešťastie ľudské hľadače vzorov fungujú tak prehnane, že vidíme vzory bez ohľadu na to, či existujú alebo nie; slabú negatívnu koreláciu si môžeme zameniť za silnú pozitívnu pri troške selektívnej pamäte.

Môžete to vidieť v pojmoch neurónových algoritmov: vytvoriť meno pre množinu vecí je ako alokovať podsieť, ktorá v nich bude hľadať vzory.

Môžete to vidieť v pojmoch chyby kompresie: veci, ktoré dostanú rovnaké meno, skončia hodené do rovnakého myšlienkového vedra, čím sa rozmažú dokopy na rovnakom bode na mape.

Alebo to môžete vidieť v pojmoch neohraničenej ľudskej schopnosti vycucať si veci z prsta a veriť im, pretože vám nikto nemôže dokázať, že sa mýlite. Akonáhle si pomenujete kategóriu, začnete si o nej vymýšľať veci. Pomenovaná vec nemusí byť vnímateľná; nemusí existovať; nemusí byť ani koherentná.

A nie, to nie je iba Japonsko: Tu na Západe bola bestsellerom kniha o diéte podľa krvnej skupiny s názvom Jedzte Správne Podľa Svojej Skupiny.

Nech sa na to pozrieme hocijako, nakresliť hranicu v priestore vecí nie je neutrálny čin. Možno čistejšie nadizajnovaná viac bayesiánska UI by dokázala myslieť na ľubovoľnú triedu a nebyť ňou ovplyvnená. Ale vy, ako človek, túto možnosť nemáte. Kategórie nie sú statické veci v kontexte ľudského mozgu; akonáhle na ne myslíte, vyvíjajú na vašu myseľ tlak. Ďalší dôvod, prečo neveriť, že si môžete definovať slovo tak, ako sa vám zachce.

*→

171. Prepašované konotácie

Včera sme videli, že v Japonsku sa krvné skupiny dostali na miesto astrológie – ak je vaša krvná skupina napríklad AB, máte byť „pohodový a vyrovnaný“.

Predstavme si, že by sme sa rozhodli vymyslieť nové slovo, „wiggin“, a *definovali* by sme, že toto slovo znamená ľudí so zelenými očami a čiernymi vlasmi...

> Zelenooký muž s čiernymi vlasmi vošiel do reštaurácie.
>
> „Ha,“ povedal Danny, pozerajúci z blízkeho stola, „videla si to? Práve sem vošiel wiggin. Prekliati wiggini. Páchajú všemožné zločiny, veru.“
>
> Jeho sestra Erda si vzdychla. „*Nevidel* si ho páchať žiaden zločin, však, Danny?“
>
> „Ani nemusím,“ povedal Danny a vybral slovník. „Pozri, píše sa to priamo tu v Oxfordskom slovníku anglického jazyka. ‚Wiggin. (1) Osoba so zelenými očami a čiernymi vlasmi.‘ Má zelené oči a čierne vlasy, je to wiggin. Nechceš sa snáď hádať s Oxfordským slovníkom anglického jazyka, alebo hej? *Podľa definície*, zelenooká čiernovlasá osoba je wiggin.“
>
> „Ale ty si ho nazval wiggin,“ povedala Erda. „To je škaredá vec povedať o niekom, koho ani nepoznáš. Nemáš žiadne indície, že si dáva priveľa kečupu na hamburgery, ani že ako dieťa používal prak na strieľanie mláďat veveričiek.“
>
> „Ale on *je* wiggin,“ povedal Danny trpezlivo. „Má zelené oči a čierne vlasy, však? Len sa pozeraj, akonáhle mu donesú hamburger, siahne po kečupe.“

Ľudská myseľ prechádza z pozorovaných vlastností k usudzovaným vlastnostiam prostredníctvom slov. Pri „všetci ľudia sú smrteľní, Sokrates je človek, preto Sokrates je smrteľný“ sú pozorované charakteristiky Sokratove šaty, reč, používanie nástrojov, a ľudský tvar vo všeobecnosti; kategória je „človek“; usudzovaná vlastnosť je otráviteľnosť bolehlavom.

Samozrejme nie je pevná hranica medzi „pozorovanými vlastnosťami“ a „usudzovanými vlastnosťami“. Ak počujete niekoho hovoriť, pravdepodobne má ľudský tvar, ak sú všetky ostatné podmienky rovnaké. Ak vidíte ľudskú postavu v tieni, *ceteris paribus*, pravdepodobne vie hovoriť.

A predsa niektoré vlastnosti majú sklon byť viac usudzované než pozorované. Skôr usúdite, že niekto je človek, a preto by horel, keby ste ho vystavili otvorenému ohňu, než aby ste odvodzovali v opačnom smere.

Ak sa pozriete do slovníka na definíciu „človeka“, pravdepodobne nájdete vlastnosti ako „inteligencia“ a „dvojnožec bez peria“ - vlastnosti, ktoré sú užitočné na rýchle omrknutie čo je a čo nie je človek – namiesto desaťtisíc konotácií od zraniteľnosti bolehlavom po prehnané sebavedomie, ktoré dokážeme usúdiť na základe toho, že niekto je človek. Prečo? Možno sú slovníky robené s úmyslom nechať vás porovnať nálepky so skupinami podobnosti, a preto sú navrhnuté, aby rýchlo izolovali zhluky v priestore vecí. Alebo možno sú veľké rozlišujúce vlastnosti tie najvýraznejšie a preto prvé napadnú redaktorovi slovníka. (Nie som si istý nakoľko si redaktori slovníkov uvedomujú, čo *v skutočnosti* robia.)

Ale výsledok je, že keď Danny vytiahne svoj slovník, aby si pozrel „wiggina“, vidí vypísané iba na prvý pohľad jasné vlastnosti, ktoré odlišujú wiggina: Zelené oči a čierne vlasy. Slovník neuvádza mnoho menších *konotácií*, ktoré sa prilepili na tento pojem, ako sú zločinecké sklony, kuchynské zvláštnosti, a niektoré nešťastné detské aktivity.

Ako sa tam tieto konotácie vôbec dostali? Možno raz existoval známy wiggin, ktorý mal tieto vlastnosti. Alebo možno si niekto náhodne vymýšľal veci a napísal o nich sériu bestsellerov (*Wiggin, Rozhovory s wigginmi, Výchova malého wiggina, Wiggin v spálni*). Možno tomu dnes veria už aj samotní

→ http://lesswrong.com/lw/nx/categorizing_has_consequences/

wiggini a správajú sa podľa toho. Akonáhle nejakých ľudí nazvete „wigginmi“, to slovo začne naberať konotácie.

Ale spomeňte si na podobenstvo o bolehlave: Ak ideme podľa logických definícií tried, nikdy nemôžeme zatriediť Sokrata ako „človeka“, dokiaľ sme nepozorovali, že je smrteľný. Kedykoľvek niekto vytiahne slovník, vo všeobecnosti sa snaží prepašovať *konotáciu*, nie skutočnú definíciu napísanú v slovníku.

Napokon, keby *jediným* významom slova „wiggin“ bolo „zelenooká čiernovlasá osoba“, prečo potom týchto ľudí nevolať iba „zelenookí čiernovlasí ľudia“? A ak sa zaujímate, či je niekto konzument kečupu, prečo sa nepýtať priamo: „Je to konzument kečupu?“ namiesto „Je to wiggin?“ (Všimnite si dosadenie podstaty na miesto symbolu.)

Aha, ale hádať sa o *skutočnej* otázke by si vyžadovalo *prácu*. Museli by ste naozaj sledovať toho wiggina, aby ste videli, či siahne po kečup. Alebo možno by ste sa museli pozrieť, či nájdete štatistiky o tom, koľko zelenookých čiernovlasých ľudí má naozaj rado kečup. V každom prípade by ste to nedokázali urobiť sediac vo svoje obývačke so zatvorenými očami. A ľudia sú leniví. Radšej budú argumentovať „podľa definície“, najmä ak si myslia, že „môžete definovať slovo, ako sa vám zachce“.

Ale samozrejme tým *skutočným* dôvodom, prečo sa starajú, či niekto je „wiggin“, je konotácia – pocit, ktorý je spojený s týmto slovom – ktorá nie je v definícii, o ktorej *tvrdia*, že ju používajú.

Predstavte si, že by Danny povedal: „Pozri, má zelené oči a čierne vlasy. Je to wiggin! Píše sa to priamo tu v slovníku! - *a preto*, má čierne vlasy. Skús proti tomu namietať, ak vieš!“

To neznie príliš triumfálne, nie? Ak by skutočná pointa argumentu naozaj *bola* zahrnutá v definícii v slovníku – keby argument naozaj *bol* logicky platný – potom by ten argument *znel* prázdno; buď by nepovedal nič nové alebo by argumentoval do kruhu.

Jedine pokus prepašovať konotácie, ktoré *nie sú* vyslovene uvedené v definícii dáva človeku pocit, že sa takýmto spôsobom dajú *získať body*.

* →

172. Argumentovanie „podľa definície“

„Toto ošklbané kura má dve nohy a nemá perie – preto, *podľa definície*, je to človek!“

Keď sa ľudia hádajú o definície, zvyčajne začnú s nejakou viditeľnou, známou, alebo aspoň často predpokladanou množinou vlastností; potom vytiahnu slovník a ukážu, že tieto vlastnosti zodpovedajú definícii v slovníku; a tak uzavrú: „Preto, *podľa definície*, ateizmus je náboženstvo!“

Lenže viditeľné, známe, často predpokladané vlastnosti sú zriedkavo skutočnou pointou debaty. Samotný fakt, že si niekto myslí, že Sokratove dve nohy sú dosť jasné, aby robili dobrý predpoklad pre argument „preto, *podľa definície*, Sokrates je človek!“ naznačuje, že dvojnohosť pravdepodobne nie je to, o čo tu *naozaj* ide – inak by poslucháč odpovedal: „To máš odkiaľ, že Sokrates je dvojnožec? Veď práve o tom sa tu od začiatku hádame!“

Existuje veru dôležitý zmysel, v ktorom sa môžeme legitímne presunúť zo zrejmých vlastností k nie celkom zrejmým. Môžete legitímne vidieť, že Sokrates má ľudský tvar a predpovedať jeho zraniteľnosť bolehlavom. Lenže tento *pravdepodobnostný* úsudok sa neopiera o definície v slovníkoch ani o bežné používanie; opiera sa o vesmír obsahujúci empirické zhluky podobných vecí.

Táto zhluková štruktúra sa nezmení podľa toho, ako si definujete svoje slová. Dokonca aj keby ste si pozreli v slovníku definíciu „človeka“ a písalo by sa tam „všetky dvojnožce bez peria okrem Sokrata“, to nezmení *skutočný* stupeň podobnosti Sokrata s nami zvyšnými dvojnožcami bez peria.

Keď teda *správne* argumentujete zo štruktúry zhluku, poviete niečo ako: „Sokrates má dve ruky, dve nohy, nos a jazyk, hovorí plynule po grécky, používa nástroje, a v každom ďalšom ohľade, čo som ho mohol pozorovať, vyzerá, že má všetky väčšie i menšie vlastnosti, ktoré charakterizujú *Homo sapiens*;

→ http://lesswrong.com/lw/ny/sneaking_in_connotations/

odhadujem teda, že má ľudskú DNA, ľudskú biochémiu, a že je zraniteľný bolehlavom, rovnako ako všetci ostatní *Homo sapiens*, na ktorých bol bolehlav klinicky testovaný na smrteľnosť.“

A predpokladajme, že ja odpoviem: „Ale ja som videl Sokrata na poli s nejakými herbológmi; myslím, že sa pokúšali pripraviť protijed. Preto *neočakávam*, že sa Sokrates zrúti, keď vypije bolehlav – bude výnimkou zo všeobecného správania sa predmetov v jeho zhluku: oni nevzali protijed, on áno.“

Teraz nemá veľký význam hádať sa, či Sokrates je „človek“ alebo nie. Rozhovor sa musí posunúť na podrobnejšiu úroveň, venovať sa *podrobnostiam*, ktoré tvoria kategóriu „človek“ - hovorilo sa o ľudskej biochémii a konkrétne o neurotoxických účinkoch koniínu.

Keby ste ďalej nástojili: „Ale Sokrates je človek a ľudia, *podľa definície*, sú smrteľní!“, potom by ste sa v skutočnosti snažili zahmliť všetko, čo vieme o Sokratovi, *okrem* faktu, že je človek – nástojili by ste, že jediná správna predpoveď je tá, ktorú by ste urobili, keby ste o Sokratovi nevedeli nič iné *okrem* toho, že je človek.

Čo je ako nástojiť na tom, že minca ukazuje hlavu alebo znak s pravdepodobnostou 50 %, pretože je to „vyvážená minca“, po tom, čo ste sa *naozaj pozreli na tú mincu* a ukazovala hlavu. Je to ako nástojiť na tom, že Frodo má desať prstov, pretože hobiti majú desať prstov, po tom, čo ste sa *pozreli na jeho ruky* a videli ste deväť prstov. Toto je samozrejme nepovolené v bayesiánskej teórii pravdepodobnosti: Nemôžete len tak odmietnuť zohľadniť novú indíciu.

A nemôžete si len tak ponechať jednu kategorizáciu a robiť odhady podľa nej, zatiaľ čo úmyselne vyhadzujete všetko ostatné, čo poznáte.

Nie každý kúsok indície robí významný rozdiel, samozrejme. Ak uvidím, že Sokrates má deväť prstov, toto viditeľne neovplyvní môj odhad jeho zraniteľnosti bolehlavom, pretože očakávam, že spôsob, ktorým Sokrates prišiel o prst, nezmenil zvyšok jeho biochémie. A to je pravda, *bez ohľadu na to, či* definícia v slovníku hovorí, že ľudia majú desať prstov. Platný úsudok sa zakladá na štruktúre zhlukov v prostredí, a na štruktúre zákonitostí biológie; *nie* na tom, čo napíše redaktor slovníka, a dokonca ani na „bežnom použití“.

Za bežných okolností, keď to robíte *správne* – *legitímnym* spôsobom – poviete iba: „Alkaloid koniín, ktorý sa nachádza v bolehlave, spôsobuje u ľudí paralýzu svalov, ktorej výsledkom je smrť zadusením.“ Alebo jednoduchšie: „Ľudia sú zraniteľní bolehlavom.“ Tak sa to zvyčajne hovorí pri *legitímnej* debate.

Kedy môže niekto cítiť potrebu *posilniť* argument dôraznou frázou „podľa definície“? (Napríklad: „Ľudia sú zraniteľní bolehlavom *podľa definície!*“) Nuž vtedy, keď sa spochybňuje usudzovaná vlastnosť – keď videli Sokrata, ako konzultuje s herbológmi – a preto hovoriaci cíti potrebu utiahnuť zverák logiky.

Keď teda vidíte takto použité „podľa definície“, zvyčajne to znamená: „Zabudni na to, čo si počul o Sokratovi konzultujúcom s herbológmi – ľudia sú *podľa definície* smrteľní!“

Ľudia cítia potrebu vmačknúť debatu na jedinú dráhu slovami: „Každé P má podľa definície vlastnosť Q!“ práve vtedy, keď vidia, a radšej by bez debaty zamietli, *dodatočné argumenty*, ktoré spochybňujú štandardné odvodenie založené na zhlukoch.

Podobne je to s argumentom: „X je *podľa definície* Y!“ Napríklad: „Ateisti veria, že Boh neexistuje; teda ateisti majú svoje presvedčenie o Bohu, pretože aj negatívne presvedčenie je presvedčenie; teda ateisti tvrdia, že majú odpovede na teologické otázky; a teda ateizmus je, *podľa definície*, náboženstvo.“

Necítili by ste potrebu povedať: „Hinduizmus, *podľa definície*, je náboženstvo!“ pretože, nuž, je samozrejmé, že Hinduizmus je náboženstvo. Nie je to iba náboženstvo „podľa definície“, ale je to *naozajstné* náboženstvo.

Ateizmus sa nepodobá na ústredných členov zhluku „náboženstvo“, takže nebyť toho, že ateizmus je náboženstvo *podľa definície*, mohli by ste si myslieť, že ateizmus *nie je* náboženstvo. Preto treba rozdrviť všetkých súperov poukázaním na to, že „ateizmus je náboženstvo“ je pravda *podľa definície*, pretože to nie je pravda podľa ničoho iného.

Čím teda hovorím: Ľudia trvajú na tom, že „X je *podľa definície* Y!“ v tých prípadoch, keď sa snažia prepašovať nejakú konotáciu z Y, ktorá nie je priamo v definícii, a X sa príliš nepodobá na zvyšných členov zhluku Y.

Počas posledných trinástich rokov som si sledoval, ako často sa tento argument používa správne verzus nesprávne – aj keď som si nerobil doslova štatistiku. Ale odhad naznačuje, že používanie frázy *podľa definície* kdekoľvek mimo matematiky je medzi najvýraznejšími varovnými signálmi chybného argumentu, aké som kedy našiel. Je zarovno s „Hitler“, „Boh“, „absolútne isté“ a „nemôžeš to dokázať“.

Táto heuristika nie je dokonalá – prvýkrát, čo som videl správne použitie mimo matematiky, to bol Richard Feynman; a odvtedy som videl ďalšie príklady. Ale pre vás bude pravdepodobne lepšie, keď jednoducho škrtnete frázu „podľa definície“ zo svojho slovníka – a *vždy*, keď budete v pokušení použiť ju zvýrazneným písmom alebo napísať za ňu výkričník. Toto je zlý nápad, *podľa definície!*

*→

173. Kde nakresliť hranicu?

Niekto k vám príde a povie:

> Dlho som uvažoval nad významom slova „Umenie“ a konečne som našiel definíciu, ktorá mi pripadá uspokojivá: „Umenie je to, čo je vytvorené za účelom vyvolať reakciu v obecenstve.“

Samotný fakt, že existuje slovo „umenie“ ešte neznamená, že ***má aj význam****, ktorý sa vznáša niekde vo vzduchoprázdne, a vy ho dokážete* ***objaviť*** *nájdením tej správnej definície.*

Pripadá vám to tak, ale nie je to tak.

Rozmýšľať, ako *definovať slovo*, znamená, že sa na problém pozeráte nesprávnym spôsobom – hľadáte tajomnú podstatu toho, čo je v skutočnosti komunikačný signál.

Existuje aj skutočná úloha, do ktorej by sa racionalista mohol legitímne pustiť, ale táto úloha nie je nájsť uspokojivú definíciu slova. Skutočnú úlohu možno hrať aj ako hru pre jedného hráča, bez rozprávania. Úlohou je zistiť, ktoré veci sa na seba podobajú – ktoré veci sa zhlukujú spolu – a niekedy, ktoré veci majú spoločnú príčinu.

Keby ste si definovali „magnetizmus“ tak, že zahŕňa blesky, zahŕňa kompasy, nezahŕňa svetlo, a zahŕňa Mesmerov „živočíšny magnetizmus“ (ktorý dnes voláme hypnóza), potom budete mať problém pýtať sa: „Ako funguje magnetizmus?“ Zlepili ste dokopy veci, ktoré k sebe nepatria, a vynechali ste iné, ktoré by ste potrebovali na doplnenie množiny. (Tento príklad je historicky uveriteľný; Mesmer bol pred Faradayom.)

Mohli by sme povedať, že magnetizmus je *zlé slovo*; hranica v priestore vecí, ktorá robí slučky a vývrtky pomedzi zhluky; rez, ktorý nerozdeľuje skutočnosť pozdĺž jej prirodzených zhybov.

Zisťovanie, kde rozrezať skutočnosť tak, aby sme krájali pozdĺž jej zhybov – *toto* je problém hoden racionalistu. Je to to, čo by sa ľudia *mali* snažiť robiť, keď sa pustia do hľadania vznášajúcej sa podstaty slova.

A nemýľte sa: je to *vedecká* úloha uvedomiť si, že potrebujete spoločné slovo na opísanie dýchania a ohňa. Nepokúšajte sa teda konzultovať s redaktormi slovníkov, lebo toto nie je ich práca.

Čo je to „umenie“? Toto slovo nemá žiadnu podstatu, ktorá by sa vznášala vo vzduchoprázdne.

Mohli by ste prísť s dlhým zoznamom vecí, ktoré voláte „umenie“ a „nie umenie“:

> *Malá fúga v G mol*: Umenie.
>
> Úder do nosa: Nie umenie.
>
> Escherova *Relativita*: Umenie.
>
> Kvet: Nie umenie.

→ http://lesswrong.com/lw/nz/arguing_by_definition/

Programovací jazyk Python: Umenie.

Kríž plávajúci v moči: Nie umenie.

Romány *Tschai* od Jacka Vancea: Umenie.

Moderné umenie: Nie umenie.

A poviete mi: „Pripadá mi intuitívne nakresliť túto hranicu, ale neviem prečo – pomôžeš mi nájsť intenziu zodpovedajúcu tejto extenzii? Vieš mi dať *jednoduchý* popis tejto hranice?“

Odpoviem: „Myslím si, že to súvisí s obdivom remesla: vynaložená práca a úžasný výsledok. Predmety, ktoré si zahrnul, majú spoločné estetické emócie, ktoré vyvolávajú, a vedomé ľudské úsilie vložené do nich so zámerom vyvolať takúto emóciu.“

Pomohol som tým, alebo som iba podvádzal v hre Tabu? Povedal by som, že zoznam, ktoré ľudské emócie sú a nie sú *estetické* je omnoho kompaktnejší než zoznam všetkého, čo je alebo nie je umenie. Môžete tieto emóce uvidieť ako svetielka na skene fMRI – hovorím to takto, aby som zdôraznil, že emócie nie sú nehmotné.

Ale samozrejme moja definícia umenia nie je tá skutočná pointa. Skutočná pointa je, že môžete nesúhlasiť s intenziou alebo s extenziou mojej definície.

Môžete povedať: „*Nie* estetická emócia je to, čo majú tieto veci spoločné; majú spoločný zámer vyvolať *ľubovoľnú* silnú emóciu, kvôli vyvolávaniu samotnému.“ To by ste nesúhlasili s mojou intenziou, mojím pokusom nakresliť krivku okolo daných bodov. Hovorili by ste: „Tvoja rovnica možno týmto bodom približne zodpovedá, ale nie je to skutočná vytvárajúca distribúcia.“

Alebo by ste mohli nesúhlasiť s mojou extenziou a povedať: „Niektoré z týchto vecí dokopy nepatria – vidím, o čo sa snažíš – ale jazyk Python by v zozname nemal byť, a moderné umenie by tam malo byť.“ (To by vás označilo za naivného ignoranta, ale aj s tým by ste mohli nesúhlasiť.) Tu je predpokladom, že naozaj existuje nejaká krivka, ktorá vytvára tento zoznam podobných a nepodobných vecí – že existuje nejaký zmysel, *aj keď ste ešte nepovedali, odkiaľ sa vzal* – ale ja som nevedomky tento zmysel stratil a zahrnul som nejaké údaje z iného generátora.

Dávno predtým než *zistíte*, čo majú spoločné elektrina a magnetizmus, môžete mať podozrenie – založené na povrchnom výzore – že „živočíšny magnetizmus“ na tento zoznam nepatrí.

Kedysi dávno si ľudia mysleli, že slovo „ryba“ zahŕňa aj delfíny. Mohli by ste sa zahrať na chytrého rečníka a povedať: „Zoznam: {losos, gupka, žralok, delfín, pstruh} je jednoducho zoznam – nemôžete povedať, že zoznam je *nesprávny*. Viem dokázať pomocou teórie množín, že tento zoznam existuje. Takže moja definícia *ryby*, čo je jednoducho tento extenzionálny zoznam, nemôže byť ‚nesprávna‘, ako tvrdíš.“

Alebo by sme sa mohli prestať hrať slaboduché hry a pripustiť, že delfín do zoznamu rýb nepatrí.

Môžete prísť so zoznamom vecí, ktoré vám *pripadajú* podobné, a pokúsiť sa odhadnúť, prečo je to tak. Ale keď nakoniec objavíte, čo mali *naozaj* spoločné, môže sa ukázať, že váš odhad bol nesprávny. Môže sa dokonca ukázať, že váš zoznam bol nesprávny.

Nemôžete sa schovať za pohodlný štít správny-podľa-definície. Aj extenzionálne aj intenzionálne definície môžu byť nesprávne, môžu nekrájať skutočnosť podĺž jej zhybov.

Kategorizovanie je úsilie odhadovať, v ktorom sa môžete pomýliť; je múdre dokázať si pripustiť, z teoretického hľadiska, že vaše definície-odhady môžu byť „pomýlené“.

* →

174. Entropia a krátke kódy

(Ak nepoznáte bayesovské odvodzovanie, toto môže byť vhodná chvíľa prečítať si Intuitívne vysvetlenie Bayesovej vety.)

→ http://lesswrong.com/lw/o0/where_to_draw_the_boundary/

Predstavte si, že máte systém X, ktorý má rovnakú pravdepodobnosť byť v hociktorom z nasledujúcich 8 stavov:

{ X1, X2, X3, X4, X5, X6, X7, X8 }

Existuje mimoriadne všadeprítomná veličina – vo fyzike, matematike a dokonca aj biológii – nazývaná *entropia*; a entropia X je 3 bity. To znamená, že sa v priemere musíte opýtať 3 otázky typu áno/nie, aby ste zistili hodnotu X. Napríklad niekto by mohol zapísať hodnoty X pomocou tohto kódu:

X1: 001 X2: 010 X3: 011 X4: 100

X5: 101 X6: 110 X7: 111 X8: 000

Takže keby som sa opýtal: „Je prvý symbol 1?“ a počul „áno“, potom sa opýtal: „Je druhý symbol 1?“ a počul „nie“, potom sa opýtal: „Je tretí symbol 1?“ a počul „nie“, vedel by som, že X je v stave 4.

Predpokladajme teraz, že systém Y má štyri možné stavy s nasledujúcimi pravdepodobnosťami:

Y1: 1/2 (50 %) Y2: 1/4 (25 %) Y3: 1/8 (12,5 %) Y4: 1/8 (12,5 %)

Potom by entropia Y bola 1,75 bitu, čo znamená, že by sme vedeli zistiť hodnotu položením v priemere 1,75 otázky typu áno/nie.

Čo znamená hovoriť o kladení jednej a troch štvrtín otázky? Predstavte si, že by sme označili stavy Y pomocou nasledujúceho kódu:

Y1: 1 Y2: 01 Y3: 001 Y4: 000

Najprv sa opýtate: „Je prvý symbol 1?“ Ak je odpoveď „áno“, skončili ste: Y je v stave 1. Toto sa stane v polovici prípadov, takže v 50 % prípadov stačí 1 otázka typu áno/nie na zistenie stavu Y.

Predpokladajme, že odpoveď je „nie“. Potom sa opýtate: „Je druhý symbol 1?“ Ak je odpoveď „áno“, skončili ste: Y je v stave 2. Y je v stave 2 s pravdepodobnosťou ¼, a vždy keď je Y v stave 2, zistíme tento fakt položením dvoch otázok typu áno/nie, takže v 25 % prípadov stačia 2 otázky na zistenie stavu Y.

Ak bola dvakrát odpoveď „nie“, opýtate sa: „Je tretí symbol 1?“ Ak „áno“, skončili ste a Y je v state 3; ak „nie“, skončili ste a Y je v stave 4. V 1/8 prípadov je Y v stave 3 a treba tri otázky; a v 1/8 prípadov je Y v stave 4 a treba tri otázky.

$(1/2 \times 1) + (1/4 \times 2) + (1/8 \times 3) + (1/8 \times 3)$

$= 0,5 + 0,5 + 0,375 + 0,375$

$= 1,75.$

Všeobecný vzorec na entropiu systému S je súčet pre všetky Si výrazu -p(Si)*log2(p(Si)).

Napríklad logaritmus (pri základe 2) čísla 1/8 je -3. Takže -(1/8 * -3) = 0,375 je príspevok stavu S4 k celkovej entropii: v 1/8 prípadov potrebujeme položiť 3 otázky.

Nemôžete vždy vymyslieť dokonalý kód pre nejaký systém, ale ak niekomu musíte povedať stav ľubovoľného množstva kópií S pomocou jedinej správy, dokážete sa k dokonalému kódu dostať ľubovoľne blízko. (Jednoduchú metódu nájdete googlením „aritmetické kódovanie“.)

Teraz sa môžete opýtať: „Prečo pre Y4 nepoužiť kód 10, namiesto 000? Neumožnilo by nám to posielať správy rýchlejšie?“

Lenže ak použijete kód 10 pre Y4, potom keď niekto odpovie „áno“ na otázku „Je prvý symbol 1?“, nebudete ešte vedieť, či je stav systému Y1 (1) alebo Y4 (10). Keby ste takto zmenili kód, celý systém sa v skutočnosti rozpadne – lebo ak počujete „1001“, neviete, či to znamená „Y4 a potom Y2“ alebo „Y1 a potom Y3“.

Ponaučenie je, že *krátke slová sú obmedzený zdroj*.

Kľúčom k vytvoreniu dobrého kódu – kódu, ktorý prenáša správy tak kompaktne, ako sa dá – je vyhradiť si krátke slová na veci, ktoré potrebujete hovoriť často, a použiť dlhšie slová na veci, ktoré nepotrebujete hovoriť tak často.

Keď dotiahnete toto umenie až po jeho hranice, dĺžka správy, ktorou potrebujete niečo opísať, zodpovedá presne alebo takmer presne jej pravdepodobnosti. Toto je formalizácia Occamovej britvy ako minimálna dĺžka popisu alebo minimálna dĺžka správy.

A tak dokonca ani *nálepky*, ktoré používame pre slová, nie sú celkom svojvoľné. Zvuky, ktoré priraďujeme našim pojmom, môžu byť lepšie alebo horšie, múdrejšie alebo hlúpejšie. Ešte aj keď neberieme do úvahy bežné použitie.

Hovorím toto všetko, pretože myšlienka „môžete mať X ako sa vám zachce“ je obrovskou prekážkou pri učení sa, ako X používať múdro. „Toto je slobodná krajina; mám právo na svoj vlastný názor,“ prekáža umeniu hľadať pravdu. „Môžem si definovať slovo, ako sa mi zachce“ prekáža umeniu krájania skutočnosti na jej zhyboch. A dokonca aj rozumne znejúce „nálepky, ktoré pridávame k slovám, sú svojvoľné“ zakrýva uvedomenie si kompaktnosti. Mimochodom, záleží aj na ľubozvučnosti – Tolkien si raz všimol, aký krásny zvuk má slovné spojenie „dvere do sklepa“; to je ten druh uvedomovania si, ktorý treba, aby ste používali jazyk ako Tolkien.

Dĺžka slov zohráva nezanedbateľnú rolu aj v kognitívnej vede jazyka:

Vezmite si slová „sklápacie kreslo“, „stolička“ a „nábytok“. Sklápacie kreslo je konkrétnejšia kategória než stolička; nábytok je všeobecnejšia kategória než stolička. Ale prevažná väčšina stoličiek má spoločné použitie – používate rovnaké pohyby, aby ste si na ne sadli, a sedíte na nich z rovnakého dôvodu (aby ste nezaťažovali svoje nohy zatiaľ čo jete alebo čítate alebo píšete alebo odpočívate). Sklápacie kreslá sa od tejto témy neoddeľujú. Na druhej strane, „nábytok“ zahŕňa veci ako postele a stoly, ktoré majú rôzny účel a vyžadujú si iné pohyby ako stoličky.

V pojmoch kognitívnej psychológie je „stolička“ *kategória na základnej úrovni*.

Ľudia majú sklon hovoriť a pravdepodobne aj myslieť na základnej úrovni kategorizácie – kresliť hranicu okolo „stoličiek“ namiesto okolo konkrétnejšej kategórie „sklápacie kreslo“ alebo všeobecnejšej kategórie „nábytok“. Ľudia častejšie hovoria: „môžeš si sadnúť na túto stoličku“ než „môžeš si sadnúť na toto sklápacie kreslo“ alebo „môžeš si sadnúť na tento nábytok“.

A nie je náhoda, že slovo „stolička“ (po anglicky: „chair“) obsahuje menej slabík než „sklápacie kreslo“ (po anglicky: „recliner“) alebo „nábytok“ (po anglicky: „furniture“). Kategórie na základnej úrovni mávajú vo všeobecnosti krátke názvy; a podstatné mená s krátkymi názvami zvyknú odkazovať na kategórie na základnej úrovni. Nie je to samozrejme dokonalé pravidlo, ale jednoznačný sklon. Časté používanie sa spája s kratšími slovami; kratšie slová sa spájajú s častým používaním.

*→

175. Vzájomná informácia a hustota v priestore vecí

Predstavte si, že máte systém X, ktorý môže byť v ľubovoľnom z 8 stavov, ktoré sú všetky rovnako pravdepodobné (relatívne k vášmu terajšiemu stavu poznania), a systém Y, ktorý môže byť v ľubovoľnom zo 4 stavov, všetkých rovnako pravdepodobných.

Entropia X, ako som včera definoval, sú 3 bity; potrebujeme položiť 3 otázky typu áno/nie, aby sme zistili presný stav X. Entropia Y, ako som včera definoval, sú 2 bity; potrebujeme položiť 2 otázky typu áno/nie, aby sme zistili presný stav Y. To môže vyzerať zrejmé, keďže $2^3 = 8$ a $2^2 = 4$, takže 3 otázky dokážu rozlíšiť 8 možností a 2 otázky dokážu rozlíšiť 4 možnosti; ale pamätajte, že keby tie možnosť neboli všetyk rovnako pravdepodobné, mohli by sme použiť chytrejší kód na odhalenie stavu Y pomocou napríklad 1,75 otázky v priemere. V tomto prípade je však *pravdepodobnostná masa* X a takisto aj Y *rozdelená rovnomerne*, takže nemôžeme použiť žiaden chytrý kód.

Aká je entropia spojeného systému (X, Y)?

Môžete byť v pokušení odpovedať: „Treba 3 otázky na zistenie stavu X, a potom 2 otázky na zistenie stavu Y, takže treba spolu 5 otázok na zistenie stavu X a Y.“

Lenže čo ak sú tieto dve premenné previazané, takže nám spoznanie stavu Y povie niečo o stave X?

→ http://lesswrong.com/lw/o1/entropy_and_short_codes/

Konkrétne, predstavme si, že X a Y sú buď obe párne alebo obe nepárne.

Teraz, ak dostanete 3-bitovú správu (opýtame sa 3 otázky) a zistíme, že X je v stave 5, vieme, že Y je v stave 1 alebo 3, ale nie v stave 2 alebo 4. Takže jediná doplňujúca otázka „je Y v stave 3?“ zodpovedaná „nie“ nám povie celý stav (X, Y): X = X5, Y = Y1. A toto sme zistili pomocou spolu 4 otázok.

Podobne, ak pomocou dvoch otázok zistíme, že Y je v stave 4, potrebujeme iba dve doplňujúce otázky, aby sme zistili, či je X v stave 2, 4, 6 alebo 8. Opäť, štyri otázky na zistenie stavu spojeného systému.

Vzájomná informácia dvoch premenných je definovaná ako rozdiel medzi entropiou spojeného systému a entropiami samostatných systémov: I(X, Y) = H(X) + H(Y) – H(X, Y).

Tu máme jeden bit vzájomnej informácie medzi dvoma systémami: Naučenie sa X nám povie jeden bit informácie o Y (zníži priestor možností zo 4 na 2, čo je 2-násobné zmenšenie objemu) a naučenie sa Y nám povie jeden bit informácie o X (zníži priestor možností z 8 na 4).

Čo ak masa pravdepodobnosti nie je rozdelená rovnomerne? Včera sme napríklad rozoberali prípad, kde Y malo pravdepodobnosť štyroch stavov ½, ¼, 1/8, 1/8. Vezmime si toto ako pravdepodobnostnú distribúciu Y, ak ho zvažujeme samostatne – keby sme videli Y a nič iné, očakávali by sme, že uvidíme toto. A predpokladajme, že premenná Z má dva stavy, 1 a 2, ktorých pravdepodobnosti sú 3/8 a 5/8.

Potom, vtedy a iba vtedy, keď je spoločná distribúcia Y a Z nasledujúca, je medzi Y a Z nulová vzájomná informácia:

Z1Y1: 3/16 Z1Y2: 3/32 Z1Y3: 3/64 Z1Y3: 3/64

Z2Y1: 5/16 Z2Y2: 5/32 Z2Y3: 5/64 Z2Y3: 5/64

Táto distribúcia sa riadi pravidlom:

$$P(Y, Z) = P(Y) \times P(Z)$$

Napríklad P(Z1Y2) = P(Z1) * P(Y2) = 3/8 × 1/4 = 3/32.

A všimnite si, že môžeme získať späť hraničné (nezávislé) pravdepodobnosti Y a Z púhym pozrením sa na spoločnú distribúciu:

P(Y1) = celková pravdepodobnosť rôznych spôsobov, ako môže nastať Y1

= P(Z1Y1) + P(Z2Y1)

= 3/16 + 5/16

= 1/2.

Takže púhym preskúmaním spoločnej distribúcie vieme určiť, či sú hraničné premenné Y a Z nezávislé; čiže, či sa spoločná distribúcia faktorizuje na súčin hraničných distribúcií; či pre každé Y a Z platí P(Y, Z) = P(Y) × P(Z).

To posledné je dôležité, pretože podľa Bayesovho pravidla:

$$P(Yi, Zj) = P(Yi) \times P(Zj)$$

$$P(Yi, Zj) / P(Zj) = P(Yi)$$

$$P(Yi|Zj) = P(Yi)$$

Po slovensky: „Keď sa dozviete Zj, váš názor na Yi je rovnaký ako bol predtým.“

Takže keď sa distribúcia faktorizuje – keď P(Y, Z) = P(Y) * P(Z) – je to *to isté* ako „Spoznanie Y nám nikdy nepovie nič o Z, a naopak.“

Z čoho ste mohli správne usúdiť, že neexistuje vzájomná informácia medzi Y a Z. Kde neexistuje vzájomná informácia, neexistujú bayesovské indície, a naopak.

Predpokladajme, že by sme v horeuvedenej distribúcii YZ považovali každú možnú kombináciu Y a Z za samostatnú udalosť – takže distribúcia YZ by mala spolu 8 možností, s uvedenými pravdepodobnosťami – a potom by sme spočítali entropiu distribúcie YZ rovnakým spôsobom, ako počítame entropiu hocijakej inej distribúcie:

3/16 × log2(3/16) + 3/32 × log2(3/32) + 3/64 × log2(3/64) + … + 5/64 × log2(5/64)

Skončili by ste s rovnakým súčtom, aký by ste dostali, keby ste samostatne spočítali entropiu Y plus entropiu Z. Neexistuje vzájomná informácia medzi týmito dvoma premennými, takže naša neistota o spojenom systéme nie je o nič menšia než naša neistota o oboch systémoch vnímaných osobitne. (Neukazujem tu výpočty, ale môžete si to spočítať sami; ani neukazujem dôkaz, že to platí vo všeobecnosti, ale môžete si vygoogliť „Shannonova entropia“ a „vzájomná informácia“.)

Čo ak sa spoločná distribúcia nefaktorizuje? Napríklad:

Z1Y1: 12/64 Z1Y2: 8/64 Z1Y3: 1/64 Z1Y4: 3/64

Z2Y1: 20/64 Z2Y2: 8/64 Z2Y3: 7/64 Z2Y4: 5/64

Ak spočítate spoločné pravdepodobnosti, aby ste dostali hraničné pravdepodobnosti, malo by vám vyjsť P(Y1) = 1/2, P(Z) = 3/8, a tak ďalej – hraničné pravdepodobnosti rovnaké ako predtým.

Ale spoločné pravdepodobnosti sa vždy nerovnajú súčinu hraničných pravdepodobností. Napríklad pravdepodobnosť P(Y1Z2) sa rovná 8/64, ale P(Z1) × P(Y2) by sa rovnako 3/8 × 1/4 = 6/64. To znamená, že pravdepodobnosť nájdenia Z1Y2 spolu je väčšia než by ste očakávali na základe pravdepodobností nájdenia Z1 alebo Y2 jednotlivo.

Z čoho následne vyplýva:

P(Z1Y2) > P(Z1)P(Y2)

P(Z1Y2)/P(Y2) > P(Z1)

P(Z1|Y2) > P(Z1)

Keďže je „nezvyčajne vysoká“ pravdepodobnosť P(Z1Y2) – definovaná ako pravdepodobnosť vyššia než by štandardne naznačovali hraničné pravdepodobnosti – vyplýva z toho, že pozorovanie Y2 je indíciou, ktorá zvyšuje pravdepodobnost Z1. A symetrickým argumentom, pozorovanie Z1 musí zvyšovať šance Y2.

Keďže existujú aspoň niektoré hodnoty Y, ktoré nám hovoria o Z (a naopak), musí medzi týmito dvoma premennými existovať vzájomná informácia; a to zistíte – som si istý, aj keď som to v skutočnosti nekontroloval – keď vám vypočítaná entropia YZ dá menej celkovej neistoty než súčet nezávislých entropií Y a Z. H(Y, Z) = H(Y) + H(Z) – I(Y, Z), kde všetky čísla sú nevyhnutne nezáporné.

(Tu odbočím, aby som podotkol, že súmernosť výrazu pre vzájomnú informáciu nám ukazuje, že Y nám musí v priemere povedať o Z rovnako veľa ako nám Z povie o Y. Ponechávam ako cvičenie pre čitateľa, aby si to uviedol do súladu so všetkým, čo sa učil na hodinách logiky o tom, že ak sú všetky havrany čierne, možnosť usudzovať Havran(x) -> Čierny(x) neznamená, že máte právo usudzovať Čierny(x) -> Havran(x). Ako odlišne vyzerajú súmerné toky pravdepodobnosti Bayesiáncov oproti ostrým zášklbom logiky – hoci to druhé je iba degenerovaným prípadom toho prvého.)

„Ale,“ pýtate sa, „ako toto všetko súvisí so správnym používaním slov?“

V kapitolách Prázdne nálepky a Nahraďte symbol podstatou sme videli techniku nahradenia slova jeho definíciou – mali sme príklad:

Všetci [smrteľný, dvojnožec, ~perie] sú smrteľní.

Sokrates je [smrteľný, dvojnožec, ~perie].

Preto, Sokrates je smrteľný.

Načo by ste potom vôbec chceli mať slovo „človek“? Prečo nepovedať iba „Sokrates je smrteľný dvojnožec bez peria“?

Pretože je užitočné mať krátke slová pre veci, s ktorými sa často stretávame. Ak je váš kód na opisovanie jednotlivých vlastností už efektívny, nebude mať výhodu použiť špeciálne slovo na ich konjunkciu – napríklad „človek“ pre „smrteľný dvojnožec bez peria“ - pokiaľ sa veci, ktoré sú smrteľné *aj* dvojnohé *aj* bez peria nenachádzajú *častejšie* než by vás hraničné pravdepodobnosti viedli očakávať.

Pri efektívnom kóde dĺžka slova zodpovedá pravdepodobnosti – preto bude kód pre Z1Y2 rovnako dlhý ako kód pre Z1 plus kód pre Y2, pokiaľ neplatí P(Z1Y2) > P(Z1) * P(Y2), lebo v takom prípade môže byť kód pre toto slovo kratší než kódy pre jeho časti.

A toto následne presne zodpovedá prípadu, kde môžeme vydedukovať niektoré vlastnosti danej veci z videnia jej iných vlastností. Musí byť väčšia než štandardná pravdepodobnosť, že dvojnohé veci bez peria budú zároveň smrteľné.

Samozrejme, slovo „človek“ v skutočnosti opisuje omnoho, omnoho viac vlastností – keď vidíte bytosť v tvare človeka, ktorá hovorí a je oblečená, môžete si o nej odvodiť celé záplavy biochemických a anatomických a kognitívnych faktov. Nahradiť slovo „človek“ opisom všetkého, čo vieme o ľuďoch, by vyžadovalo, aby sme strávili neprimerané množstvo času rozprávaním. Ale toto platí *iba* vďaka tomu, že rozprávajúci dvojnožec bez peria má omnoho väčšiu než štandardnú pravdepodobnosť byť otráviteľný bolehlavom alebo mať široké nechty alebo mať prehnané sebavedomie.

Mať pre nejakú vec slovo, namiesto iba vymenovania jej vlastností, je kompaktnejší kód práve v tých prípadoch, kde dokážeme usúdiť niektoré z týchto vlastností z iných vlastností. (Možno s výnimkou veľmi primitívnych slov ako „červená“, ktoré používame na vyslanie celkom nekomprimovaného popisu našich zmyslových vnemov. Ale keď sa stretnete s chrobákom alebo čo len kameňom, už máte do činenia s nejednoduchou zbierkou vlastností, vysoko nad primitívnou úrovňou.)

Mať teda slovo „wiggin“ pre zelenookých čiernovlasých ľudí je užitočnejšie než iba povedať „zelenooká čiernovlasá osoba“ práve vtedy, keď:

1. Zelenookí ľudia majú väčšiu než priemernú pravdepodobnosť byť čiernovlasí (a naopak), čo znamená, že môžeme pravdepodobnostne usudzovať na zelené oči z čiernych vlasov a naopak; *alebo*
2. Wiggini majú spoločné ďalšie vlastnosti, na ktoré možno usudzovať s pravdepodobnosťou väčšou než štandardnou. V tomto prípade musíme osobitne pozorovať zelené oči a čierne vlasy; ale keď už sme pozorovali obe tieto vlastnosti jednotlivo, môžeme pravdepodobnostne odvodiť iné vlastnosti (napríklad chuť na kečup).

Mohli by sme priam považovať samotný akt definovala slova ako prísľub takéhoto účinku. Povedať niekomu: „Definujem slovo ‚wiggin‘ ako osobu so zelenými očami a čiernymi vlasmi,“ podľa Griceovských implikácií tvrdí, že slovo „wiggin“ vám nejako pomôže usudzovať / skracovať svoje správy.

Ak zelené oči a čierne vlasy nemajú väčšiu než štandardnú pravdepodobnosť vyskytovať sa spolu, ani sa žiadna iná vlastnosť nevyskytuje v ich prítomnosti s väčšiu než štandardnou pravdepodobnosťou, potom samotné slovo „wiggin“ je lož: Toto slovo tvrdí, že určitých ľudí sa oplatí rozlišovať ako skupinu, ale nie je to tak.

V tomto prípade slovo „wiggin“ nepomáha opisovať skutočnosť kompaktnejšie – nie je definované niekým, kto posiela najkratšiu správu – nemá miesto v najjednoduchšom vysvetlení. To znamená, že slovo „wiggin“ vám nijako nepomôže pri robení žiadnych bayesovských odvodení. Aj keby ste to slovo neoznačili za lož, rozhodne je to chyba.

A správny spôsob, ako rezať skutočnosť pozdĺž jej záhybov je kresliť svoje hranice okolo koncentrácií nezvyčajne vysokej hustoty pravdepodobnosti v priestore vecí.

* →

176. Superexponenciálny priestor pojmov a jednoduché slová

Možno si myslíte, že priestor vecí je pomerne veľký priestor. Omnoho väčší než skutočnosť, pretože kde skutočnosť obsahuje iba veci, ktoré naozaj existujú, priestor vecí obsahuje všetko, čo by *mohlo* existovať.

→ http://lesswrong.com/lw/o2/mutual_information_and_density_in_thingspace/

V skutočnosti, ako som „definoval“ priestor vecí, že má rozmer pre každú možnú vlastnosť – vrátane korelujúcich vlastností ako hustota a objem a hmotnosť – je priestor vecí asi príliš zle definovaný na to, aby mal niečo, čo by sa dalo nazvať *veľkosťou*. Ale *aj tak* je dôležité vedieť si priestor vecí predstaviť. Iste nikto nedokáže *naozaj* pochopiť kŕdeľ vrabcov, pokiaľ vidí iba obláčik mávajúcich čvirikajúcich vecí, namiesto zhluku bodov v priestore vecí.

Ale akýkoľvek veľký by priestor vecí mohol byť, nesiaha ani po päty priestoru pojmov.

„Pojem“ v strojovom učení znamená pravidlo, ktoré zahŕňa alebo vylučuje príklady. Ak vidíte údaje { 2:áno, 3:nie, 14:áno, 23:nie, 8:áno, 9:nie }, potom možno uhádnete, že tento pojem bol „párne čísla“. Existuje pomerne veľa literatúry (ako sa dalo čakať) o tom, ako sa učiť pojmy na základe údajov... ak dostanete náhodné príklady, ak dostanete vybrané príklady... ak sa v triedení môžu vyskytovať chyby... a najmä, ak máte rozličné priestory možných pravidiel.

Predstavme si napríklad, že sa chceme naučiť pojem „dobrý deň na hranie tenisu“. Možné vlastnosti Dní sú:

Obloha: { Slnečno, Zamračené, Dážď }

TeplotaVzduchu: { Teplo, Zima }

Vlhkosť: { Normálna, Vysoká }

Vietor: { Silný, Slabý }

Potom dostaneme nasledujúce údaje, kde + znamená príklad daného pojmu, a – znamená, že toto nie je príklad daného pojmu:

\+ Obloha: Slnečno; TeplotaVzduchu: Teplo; Vlhkosť: Vysoká; Vietor: Silný.

\- Obloha: Dážď; TeplotaVzduchu: Zima; Vlhkosť: Vysoká; Vietor: Silný.

\+ Obloha: Slnečno; TeplotaVzduchu: Teplo; Vlhkosť: Vysoká; Vietor: Slabý.

Čo by z tohto mohol algoritmus usúdiť?

Učiaci sa stroj by si mohol predstaviť *jeden* pojem, ktorý je v súlade s týmito údajmi:

Obloha: ?; TeplotaVzduchu: Teplo; Vlhkosť: Vysoká; Vietor: ?

V tomto formáte, aby sme zistili, či tento pojem zahŕňa alebo vylučuje nejaký príklad, porovnávame prvok za prvkom: ? akceptuje hocičo, ale konkrétna hodnota akceptuje iba túto konkrétnu hodnotu.

Takže horeuvedený pojem bude akceptovať iba Dni, kde TeplotaVzduchu = Teplo a Vlhkosť = Vysoká, ale Obloha a Vietor môžu mať ľubovoľnú hodnotu. Toto je v súlade s negatívnymi aj pozitívnymi klasifikáciami doterajších údajov – ale nie je to *jediný* takýto pojem.

Reprezentáciu horeuvedeného pojmu si môžete zjednodušiť na { ?, Teplo, Vysoká, ? }

Bez zachádzania do podrobností, klasický algoritmus by znel:

- Udržuj si množinu najvšeobecnejších hypotéz, ktoré sú v súlade s údajmi – tie, ktoré pozitívne zatrieďujú čo najvia príkladov, ale pritom sú v súlade s faktmi.
- Udržuj si ďalšiu skupinu čo najkonkrétnejších hypotéz, ktoré sú v súlade s údajmi – tie, ktoré negatívne zatrieďujú čo najviac príkladov, ale pritom sú v súlade s faktmi.
- Vždy, keď vidíme nový negatívny príklad, posilníme všetky tie najvšeobecnejšie hypotézy čo najmenej, aby nová množina bola opäť taká všeobecná, ako sa len dá, ale v súlade s faktmi.
- Vždy, keď vidíme nový pozitívny príklad, uvoľníme všetky tie najkonkrétnejšie hypotézy čo najmenej, aby nová množina bola opäť taká konkrétna, ako sa len dá, ale v súlade s faktmi.
- Pokračujeme, dokiaľ nám nedostane iba jediná hypotéza. Toto bude odpoveď, *ak* daný pojem vôbec bol v našom priestore hypotéz.

V horeuvedenom prípade by množina najvšeobecnejších hypotéz bola { ?, Teplo, ?, ? } a { Slnečno, ?, ?, ? }, zatiaľ čo množina najkonkrétnejších hypotéz má jediný prvok { Slnečno, Teplo, Vysoká, ? }.

Ľubovoľný iný nájdený pojem, ktorý zodpovedá daným údajom, bude striktne konkrétnejší než tie najvšeobecnejšie hypotézy, a striktne konkrétnejší než tá najvšeobecnejšia hypotéza.

(Ďalšie informácie nájdete v knihe *Strojové učenie* od Toma Mitchella, odkiaľ som si upravil tento príklad.[171])

Možno ste si všimli, že horeuvedený formát *nedokáže* reprezentovať všetky možné pojmy. Napríklad: „Hraj tenis, keď je obloha slnečná *alebo* vzduch teplý.“ Toto je v súlade so všetkými danými údajmi, ale v horeuvedenej reprezentácii pojmov žiadna štvorica hodnôt nepopisuje toto pravidlo.

Je jasné, že naše strojové učenie nie je veľmi všeobecné. Prečo mu nedovoliť reprezentovať *všetky možné* pojmy, aby sa dokázalo učiť s najväčšou možnou pružnosťou?

Dni sa skladajú zo štyroch premenných, jedna z ich má 3 hodnoty a tri z nich majú po 2 hodnoty. Môžeme teda vidieť $3 \times 2 \times 2 \times 2 = 24$ možných Dní.

Formát, ktorý sme použili na reprezentáciu pojmov nám dovoľuje vyžadovať pre každú premennú ľubovoľnú z týchto hodnôt, alebo ponechať danú premennú otvorenú. V tejto reprezentácii teda máme $4 \times 3 \times 3 \times 3 = 108$ pojmov. Aby nám fungoval najvšeobecnejší/najkonkrétnejší algoritmus, potrebujeme začať s najkonkrétnejšou hypotézou: „žiaden príklad nie je nikdy pozitívne klasifikovaný“. Keď to pridáme, máme dokopy 109 pojmov.

Je podozrivé, že je viac možných pojmov než je možných Dní? Iste nie: Napokon, každý pojem môžeme vnímať ako *množinu* Dní. Pojem možno vnímať ako množinu dní, ktoré klasifikuje pozitívne, alebo izomorfne, ako množinu dní, ktoré klasifikuje negatívne.

Takže priestor *všetkých možných* pojmov, ktoré klasifikujú Dni je množina všetkých možných množín Dní, ktorej veľkosť je $2^{24} = 16\,777\,216$.

Tento úplný priestor zahŕňa všetky pojmy, o ktorých sme doteraz hovorili. Ale zahŕňa aj pojmy ako „Pozitívne klasifikuj iba príklady { Slnečno, Teplo, Vysoká, Silný } a { Slnečno, Teplo, Vysoká, Slabý } a odmietni všetky ostatné“ alebo „Negatívne klasifikuj iba jediný príklad { Dážď, Zima, Vysoká, Silný } a prijmi všetko ostatné.“ Zahŕňa javy, ktoré nemajú žiadnu kompaktnú reprezentáciu; sú to iba zoznamy toho, čo je a čo nie je dovolené.

Toto je problém, keď sa snažíme zostrojiť „plne všeobecný“ induktívny učiaci sa stroj. Nedokáže sa naučiť pojmy, dokiaľ nevidel všetky možné príklady v priestore inštancií.

Ak pridáme k Dňom ďalšie vlastnosti – napríklad teplota vody alebo predpoveď na zajtra – počet možných dní porastie exponenciálne s počtom vlastností. Ale to nie je problém v našom obmedzenom priestore pojmov, pretože dokážeme zachytiť veľký priestor pomocou logaritmického množstva príkladov.

Povedzme, že sme dňom pridali vlastnosť Voda: { Teplá, Studená }, čo nám dáva 48 možných Dní a 325 možných pojmov. Povedzme, že každý Deň, ktorý vidíme, zvyčajne asi polovica momentálne dôveryhodných pojmov klasifikuje kladne a druhá polovica záporne. Keď sa teda dozvieme skutočnú klasifikáciu tohto príkladu, rozdelí to priestor kompatibilných pojmov napoly. Možno nám teda stačí 9 príkladov ($2^9 = 512$) na zúženie 325 možných pojmov na jeden.

Aj keby Dni mali štyridsať binárnych vlastností, stále by si to malo vyžadovať iba zvládnuteľné množstvo údajov, aby sme zúžili možné pojmy na jeden. 64 príkladov, keby každý príklad klasifikovala pozitívne polovica zostávajúcich pojmov. Samozrejme *za predpokladu*, že *skutočné* pravidlo vôbec dokážeme reprezentovať!

Ak chcete myslieť na všetky možnosti, nuž, veľa šťastia. Priestor *všetkých možných* pojmov rastie *super*exponenciálne s počtom vlastností.

V čase, keď hovoríte o údajoch so štyridsiatimi binárnymi vlastnosťami, počet možných príkladov je vyše bilióna – ale počet možných *pojmov* je vyše dve na bilióntu. Aby ste zúžili tento *super*exponenciálny priestor pojmov, potrebovali by ste vidieť vyše bilióna príkladov predtým než by ste mohli povedať, čo je In a čo je Out. Potrebovali by ste vlastne vidieť všetky možné príklady.

Pripomínam, to je pri štyridsiatich binárnych vlastnostiach. 40 bitov, alebo 5 bajtov, ktoré majú byť klasifikované jednoduchým „áno“ alebo „nie“. Zo 40 bitov vyplýva 2^{40} možných príkladov a $2^{(240)}$ možných pojmov, ktoré tieto príklady klasifikujú ako pozitívne alebo negatívne.

171 Tom M. Mitchell, *Machine Learning* (McGraw-Hill Science/Engineering/Math, 1997).

Preto tu, v skutočnom svete, kde na opis predmetov treba viac než 5 bajtov, *a* kde nemáme k dispozícii bilión príkladov, *a* kde je v cvičných dátach šum, *myslíme* iba na *veľmi pravidelné* pojmy. Ľudská myseľ – dokonca ani celý viditeľný vesmír – nie je dosť veľká na to, aby zvážila všetky možné hypotézy.

Z tohto pohľadu sa učenie nielenže *spolieha na* induktívne skreslenie, ale sa *takmer celé* skladá z induktívneho skreslenia – keď porovnávate počet pojmov odmietnutých *a priori* s tými, ktoré ste odmietli na základe púhych indícií.

Ale ako toto (pýtate sa) súvisí so správnym používaním slov?

Je to celý dôvod, preto slová majú intenzie rovnako ako extenzie.

Včerajší článok som uzavrel:

> Spôsob, ako krájať skutočnosť pozdĺž jej zhybov, je kresliť hranice okolo koncentrácií nezvyčajne vysokej hustoty pravdepodobnosti.

Úmyselne som z tohto (mierne upraveného) výroku vynechal jednu kľúčovú vec, pretože som ju doteraz nevedel vysvetliť. Lepšie tvrdenie by bolo:

> Spôsob, ako krájať skuočnosť pozdĺž jej zhybov, je kresliť *jednoduché* hranice okolo koncentrácií nezvyčajne vysokej hustoty pravdepodobnosti v priestore vecí.

V opačnom prípade by ste si priestor vecí mohli rozdeliť umelo. Mohli by ste si vytvoriť veľmi čudné nespojité hranice, ktoré by zhromažďovali všetky pozorované príklady, príklady, ktoré by sa nedali opísať kratšou správou než samotné vaše pozorovania, a povedať: „Toto je to, čo som videl v minulosti a preto očakávam, že toho uvidím viac v budúcnosti."

V skutočnom svete sa nič nad úrovňou molekúl neopakuje *presne*. Sokrates má podobný tvar ako mnoho ďalších ľudí, ktorí sú zraniteľní bolehlavom, ale nemá presne *rovnaký* tvar ako oni. Takže váš odhad, že Sokrates je „človek" závisí od kreslenia *jednoduchých* hraníc okolo ľudského zhluku v priestore vecí. Namiesto „veci vyzerajúce presne ako [5-megabajtový opis tvaru číslo 1] a [veľa ďalších vlastností], *alebo* presne ako [5-megabajtový opis tvaru číslo 2] a [veľa ďalších vlastností] … sú ľudia".

Ak okolo svojich vnemov nekreslíte *jednoduché* hranice, nemôžete z nich usudzovať. Preto sa pokúšate opísať „umenie" pomocou intenzionálnej definície ako „to, čo má v úmysle vyvolať nejakú zložitú emóciu, za účelom samotného jej vyvolania" namiesto ukazovania na dlhé zoznamy vecí, ktoré sú alebo nie sú umenie.

V skutočnosti samotná veta „ako krájať skutočnosť pozdĺž jej zhybov" je trochu typu vajce a sliepka: Nemôžete určiť *hustotu* skutočných pozorovaní, dokiaľ ste neurobili aspoň trochu krájania. A pravdepodobnostná distribúcia vychádza z nakreslenia hraníc, nie naopak – keby ste už *mali* pravdepodobnostnú distribúciu, mali by ste všetko, čo potrebujete na usudzovanie, takže načo by ste sa obťažovali kresliť hranice?

A toto ukazuje ďalší – áno, ešte ďalší – dôvod byť podozrievavý voči tvrdeniu, že „môžete definovať slovo ako sa vám zachce". Keď si vezmete superexponenciálnu veľkosť priestoru pojmov, začne byť jasné, že vybrať si z jeho na uváženie jeden konkrétny pojem je činom nie malej drzosti – nielen pre nás, ale pre ľubovoľnú myseľ s obmedzenou výpočtovou silou.

Predložiť nám slovo „wiggin" definované ako „čiernovlasá zelenooká osoba" bez nejakého dôvodu prečo vyzdvihnúť *tento konkrétny pojem* na úroveň našej vedomej pozornosti, je skoro ako keď detektív povie: „Nuž, nemám ani najmenšiu smietku dôvodov jedným ani druhým smerom ohľadom toho, kto zavraždil tieto siroty... podotýkam, že ani len intuíciu... ale zamysleli sme sa už nad tým, či by John Q. Wiffleheim z Norkle Rd 1234 nemohol byť podozrivý?"

* →

→ http://lesswrong.com/lw/o3/superexponential_conceptspace_and_simple_words/

177. Podmienená nezávislosť a naivný Bayes

Hovoril som už o vzájomnej informácii medzi X a Y, zapísanej I(X, Y), čo je rozdiel medzi entropiou spoločnej pravdepodobnostnej distribúcie, H(X, Y) a entropiami hraničných distribúcií, H(X) + H(Y).

Dal som príklad premennej X, ktorá má osem stavov 1...8, ktoré sú všetky rovnako pravdepodobné, ak sme ešte nevideli žiadne indície; a premennej Y, ktorej stavy 1...4 sú všetky rovnako pravdepodobné, ak sme ešte nevideli žiadne indície. Ak potom vypočítame hraničné entropie H(X) a H(Y), zistíme, že X má 3 bity entropie a Y má 2 bity.

Vieme však aj, že X a Y sú buď obe párne alebo obe nepárne; a to je všetko, čo ich vzťahu vieme. Spoločná distribúcia (X, Y) má teda iba 16 možných stavov, všetky rovnako pravdepodobné, čiže spoločná entropia je 4 bity. To je 1 bit entropie menej v porovnaní s 5 bitmi entropie, keby X a Y boli nezávislé. Táto chýbajúca entropia je vzájomná informácia – informácia, ktorú nám X hovorí o Y a naopak, takže nie sme až takí neistí ohľadom jedného, keď sme sa dozvedeli druhé.

Predpokladajme však, že existuje tretia premenná Z. Z má dva stavy, „párne“ a „nepárne“, perfektne korelované k párnosti alebo nepárnosti (X, Y). V skutočnosti môžeme predpokladať, že Z je jednoducho otázka: „Sú X a Y párne alebo nepárne?“

Ak nemáme žiadne indície o X a Y, potom samotné Z nevyhnutne má 1 bit entropie danej informácie. Je 1 bit vzájomnej informácie medzi Z a X, a 1 bit vzájomnej informácie medzi Z a Y. A, ako sme povedali predtým, 1 bit vzájomnej informácie medzi X a Y. Koľko entropie má teda celý systém (X, Y, Z)? Naivne by ste mohli očakávať, že:

H(X, Y, Z) = H(X) + H(Y) + H(Z) – I(Z, X) – I(Z, Y) – I(X, Y)

ale ukáže sa, že to tak nie je.

Spojený systém (X, Y, Z) má iba 16 možných stavov – keďže Z je iba otázka „Sú X a Y párne alebo nepárne?“ - takže H(X, Y, Z) = 4 bity.

Ale keby ste počítali podľa uvedeného vzorca, dostali by ste:

(3 + 2 + 1 – 1 – 1 – 1) bity = 3 bity = ZLE!

Prečo? Pretože ak máte vzájomnú informáciu medzi X a Z, a vzájomnú informáciu medzi X a Y, môže to zahŕňať niečo z *tej istej* vzájomnej informácie, ktorá existuje medzi X a Y. Napríklad v tomto prípade nám poznanie, že X je párne, hovorí, že Z je párne, a poznanie, že Z je párne, nám hovorí, že Y je párne, ale túto istú informáciu by nám povedalo X o Y. Niečo z našich vedomostí sme započítali dvojmo a preto nám vyšlo príliš málo entropie.

Správny vzorec je (myslím si):

H(X, Y, Z) = H(X) + H(Y) + H(Z) – I(Z, X) – I(Z, Y) – I(X, Y | Z)

Pričom posledný výraz I(X, Y | Z) znamená „informácia, ktorú nám X povie o Y, za predpokladu, že už poznáme Z“. V tomto prípade nám X nepovie o Y nič, pokiaľ už poznáme Z, takže tento člen vychádza ako nula – a rovnica dáva správnu odpoveď. Tak, nie je to pekné?

„Nie,“ odpoviete správne, „pretože si mi nepovedal, ako sa *počíta* I(X, Y | Z), iba si mi dal slovný argument, že by to mala byť nula.“

Výraz I(X, Y | Z) vypočítame presne tak, ako by ste očakávali. I(X, Y) = H(X) + H(Y) – H(X, Y), preto:

I(X, Y | Z) = H(X | Z) + H(Y | Z) – H(X, Y | Z)

A teraz asi chcete vedieť, ako sa počíta podmienená entropia? Nuž, *pôvodný* vzorec pre entropiu je:

H(S) = Suma i: p(Si) × -log2(p(Si))

Keby sme sa potom dozvedeli nový fakt Z0, naša zostávajúca neistota ohľadom S by bola:

H(S | Z0) = Suma i: p(Si | Z0) × -log2(p(Si | Z0))

Ak sa teda ideme dozvedieť nový fakt Z, ale zatiaľ ešte nevieme, ktoré Z to bude, mali by sme v priemere očakávať, takúto výslednú neistotu ohľadom S:

H(S | Z) = Suma j: (p(Zj) × Suma i: p(Si | Zj) × -log2(p(Si | Zj)))

A takto sa počítajú podmienené entropie; z ktorých potom môžeme dostať podmienenú vzájomnú informáciu.

Existujú tu *všelijaké* pomocné vety ako:

H(X | Y) = H(X, Y) – H(Y)

a

ak I(X, Z) = 0 a I(X, Y | Z) = 0, potom I(X, Y) = 0

ale tam nebudem zachádzať.

„Ale“, pýtate sa, „čo má *toto* spoločné s podstatou slov a ich skrytou bayesovskou štruktúrou?“

Som *nevýslovne* rád, že ste položili túto otázku, pretože som vám to mal v úmysle povedať, či už chcete alebo nie. Ale najprv je tu ešte pár úvodných poznámok.

Budete si pamätať – áno, *budete* si pamätať – že existuje vzťah medzi vzájomnou informáciu a bayesovskou indíciou. Vzájomná informácia je kladná vtedy a iba vtedy, keď sa pravdepodobnosť aspoň niektorých spoločných udalostí P(x, y) nerovná súčinu pravdepodobností samostatných udalostí P(x) × P(y). A toto je zase presne ekvivalentné podmienke, že existuje bayesovská indícia medzi x a y:

I(X, Y) > 0 =>

P(x, y) != P(x) × P(y)

P(x, y) / P(y) != P(x)

P(x | y) != P(x)

Ak používate podmienku Z, prispôsobíte tomu celé odvodenie:

I(X, Y | Z) > 0 =>

P(x, y | z) != P(x | z) × P(y, z)

P(x, y | z) / P(y | z) != P(x | z)

(P(x, y, z) / P(z)) / (P(y, z) / P(z)) != P(x | z)

P(x, y, z) / P(y, z) != P(x | z)

P(x | y, z) != P(x | z)

Pričom posledný riadok čítame: „Aj keď už vieme Z, dozvedieť sa Y stále mení náš názor na X.“

Naopak, v našom pôvodnom prípade, kde Z bolo „párne“ alebo „nepárne“, Z oddeľovalo X od Y – čiže ak sme vedeli, že Z je „párne“, potom dozvedieť sa, že Y je v stave 4 nám už nepovedalo *nič viac* o tom, či X je 2, 4, 6 alebo 8. Alebo ak sme vedeli, že Z je „nepárne“, potom dozvedieť sa, že X je 5 nám už nepovedalo nič viac o tom, či Y je 1 alebo 3. Dozvedieť sa Z spôsobilo, že X a Y sa stali *podmienene nezávislé*.

Podmienená nezávislosť je mimoriadne dôležitý pojem v teórii pravdepodobnosti – aby som citoval iba jeden príklad, bez podmienenej nezávislosti by vesmír nemal žiadnu štruktúru.

Dnes však mám v úmysle hovoriť iba o jednom konkrétnom druhu podmienenej nezávislosti – o prípade strednej premennej, ktorá oddeľuje ostatné premenné okolo nej, ako telo s chápadlami.

Majme päť premenných U, V, W, X, Y; a ďalej predpokladajme, že pre každú dvojich týchto premenných je jedna premenná indíciou o druhej. Ak si vyberiete napríklad U a W, potom zistenie, že U = U1, vám môže povedať niečo, čo ste predtým nevedeli o pravdepodobnosti, že W = W1.

Nezvládnuteľná inferenčná zmes? Divoké indície? Nie nutne.

Možno U je „hovorí jazykom“, V je „dve ruky a desať prstov“, W je „nosí šaty“, X je „otráviteľný bolehlavom“, a Y je „červená krv“. Keď teraz stretnete nejakú vec zo skutočného sveta, ktorá by mohla byť jablko a mohla by byť kameň, a dozviete sa, že táto vec hovorí po čínsky, budete mať sklon pripísať omnoho väčšiu pravdepodobnosť tomu, že nosí šaty; a ak sa dozviete, že táto vec nie je otráviteľná bolehlavom, budete pripisovať trochu nižšiu pravdepodobnosť tomu, že má červenú krv.

Niektoré z týchto pravidiel sú silnejšie než iné. Je tu napríklad Fred, ktorému chýba prst vďaka nehode pri sopke, a je tu bábätko Barney, ktoré ešte nerozpráva, a je tu robot Irving, ktorý vypisuje vety,

ale nemá krv. Ak sa teda dozvieme, že nejaká vec nenosí šaty, to neoddeľuje všetko, čo nám schopnosť reči môže povedať o farbe krvi. Ak táto vec nenosí šaty, ale *rozpráva*, môže to byť nahá Nellie.

To robí tento príklad omnho zaujímavejším než povedzme päť celočíselných premenných, ktoré sú buď všetky párne alebo všetky nepárne, ale inak medzi nimi nie je súvislosť. V takom prípade by poznanie *ľubovoľnej* jednej premennej oddelilo všetko, čo nám poznanie druhej premennej môže povedať o tretej premennej.

To však máme závislosti, ktoré nezmiznú akonáhle sa dozvieme iba jednu premennú, ako nám ukazuje prípad nahej Nellie. Je to teda nezvládnuteľná inferenčná nepohoda?

Nebojte sa! Môže totiž existovať nejaká *šiesta* premenná Z, ktorá, keby sme ju poznali, by naozaj *dokázala* od seba oddeliť každú dvojicu premenných. Mohla by byť taká premenná Z – aj keby sme ju mali *vytvoriť* namiesto priameho pozorovania – že:

$$p(u \mid v, w, x, y, z) = p(u \mid z)$$

$$p(v \mid u, w, x, y, z) = p(v \mid z)$$

$$p(w \mid u, v, x, y, z) = p(w \mid z)$$

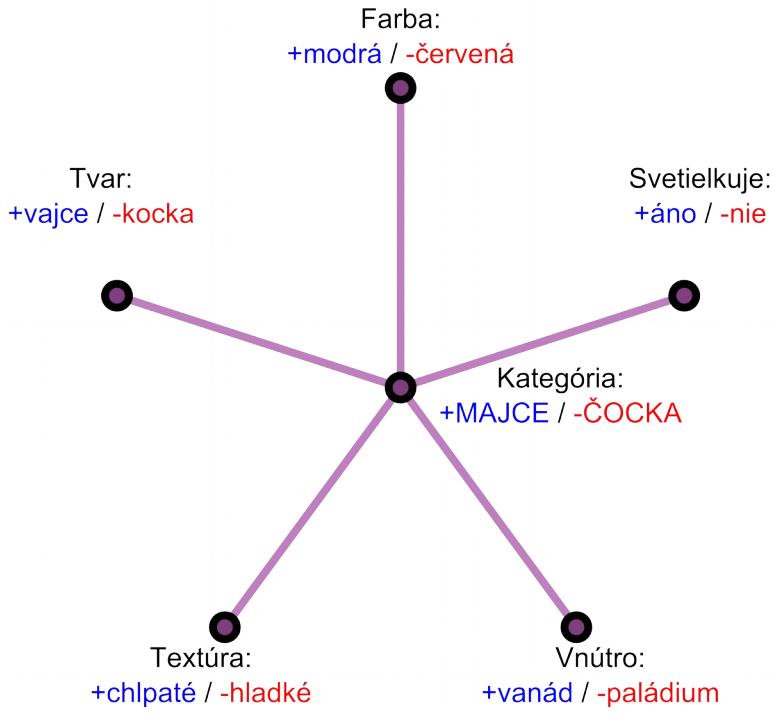
…

Možno, *keby* nejaká vec bola „človek“, potom by pravdepodobnosti, že hovorí, nosí šaty, a má štandardný počet prstov, boli všetky nezávislé. Fredovi možno chýba prst – ale nemá o nič väčšiu pravdepodobnosť než hocikto iný, že bude nudistom; nahá Nellie nikdy nenosí šaty, ale to nerobí o nič menej pravdepodobným, že rozpráva; a bábätko Barney zatiaľ nerozpráva, ale nechýbajú mu žiadne končatiny.

Toto sa nazýva metóda „naivného Bayesa“, pretože to zvyčajne nie je celkom pravda, ale *predpokladanie*, že to pravda je, môže zjednodušiť vaše výpočty. Nebudete si držať oddelené záznamy o vplyve šiat na schopnosť reči pri danom počte prstov. Iba použijeme všetky pozorované informácie na zaznamenanie pravdepodobnosti, že táto vec je človek (alebo niečo iné, napríklad šimpanz alebo robot) a potom použijeme naše názory na centrálnu triedu, aby sme predpovedali všetko, čo sme zatiaľ nepozorovali, napríklad zraniteľnosť bolehlavom.

Ľubovoľné pozorovanie U, V, W, X a Y funguje iba ako indícia pre premennú centrálnej triedy Z, a potom použijeme výslednú distribúciu Z na potrebné predpovede ohľadom zatiaľ nepozorovaných premenných medzi U, V, W, X a Y.

Znie vám to povedome? Malo by; pozrite si obrázok 117.1.



Obrázok 117.1: Sieť 2

V skutočnosti, ak použijete správny typ jednotiek neurónovej siete, táto „neurónová sieť“ skončí ako *presný, matematický* ekvivalent naivného Bayesa. Centrálna jednotka potrebuje správnu logistickú hranicu – reakciu podľa krivky tvaru S – a váhy vstupov potrebujú zodpovedať logaritmom pomerov podmienených pravdepodobnosť, atď. V skutočnosti je rozumným predpokladom, že toto je jeden z dôvodov, prečo logistická reakcia často tak dobre funguje v neurónových sieťach – umožňuje to algoritmu prepašovať trochu bayesovského uvažovania, kým sa jeho dizajnéri nepozerajú.

Keď vám niekto ukazuje algoritmus, ktorý nazýva „neurónová sieť“, oblepený pôsobivými slovíčkami ako „obyčajný“ a „emergentný“, a hrdo pritom vyhlasuje, že nemá poňatie, ako táto sieť funguje po naučení – nepredpokladajte, že ich malý algoritmus UI *naozaj* je za hranicami logiky. Ak bude táto ukážka náhodnosti fungovať, ukáže sa, že má bayesovskú štruktúru; možno bude dokonca presne ekvivalentná algoritmu toho typu, ktoré voláme „bayesovské“.

Aj keď na povrchu *nevyzerá* bayesovsky.

A potom už *viete*, že bayesovci začnú vysvetľovať, ako presne tento algoritmus funguje, ktoré základné predpoklady odráža, ktoré pravidelnosti prostredia využíva, kde funguje a kde zlyháva, a dokonca pripoja pochopiteľné významy váham naučenej siete.

To je sklamanie, čo?

*→

178. Slová ako rúčky myšlienkových štetcov

Predstavte si, že vám poviem: „Stala sa čudná vec: Lampy v tomto hoteli majú trojuholníkové žiarovky.“

Možno ste si ich predstavili, možno nie – ak nie, urobte to teraz – ako vo vašej predstave vyzerá taká „trojuholníková žiarovka“?

Vo vašej predstave, má toto sklo hrany ostré alebo oblé?

Keď sa v mojej mysli objavilo slovné spojenie „trojuholníkové žiarovky“ - nie, nemali ich v hoteli – pokiaľ možno dôverovať mojej introspekcii, najprv som videl pyramídovú žiarovku s ostrými hranami, potom (takmer ihneď) sa hrany zaoblili, a moja myseľ si vytvorila slučku fluorescenčnej žiarovky v tvare zaobleného trojuholníka ako alternatívu.

Pokiaľ môžem povedať, neboli v tom zahrnuté žiadne vedomé/slovné myšlienky – iba bezslovné reflexné cúvnutie pred predstavou ostrého skla, a tento problém dizajnu sa vyriešil skôr než som naň vôbec stihol slovne pomyslieť.

Verte tomu alebo nie, niekoľko desaťročí sa viedla vážna debata o tom, či ľudia *naozaj* majú v hlave myšlienkové obrazy – či niekde naozaj majú *obrázok* stoličky – alebo či si ľudia iba naivne *myslia*, že majú myšlienkové obrazy (pretože ich zavádza „introspekcia“, veľmi zlá zakázaná aktivita), zatiaľ čo v skutočnosti majú vo svojom mozgu aktívnu iba malú nálepku „stolička“, ako symbol LISPu.

Veľmi sa snažím nehovoriť veci ako: „Aká kolosálna blbosť,“ pretože vždy treba vziať do úvahy efekt spätného pohľadu, ale: aká kolosálna blbosť.

Myslím si, že táto akademická paradigma bola pomäteným dedičstvom behaviorizmu, ktorý popieral existenciu ľudských myšlienok, a snažil sa vysvetliť všetky ľudské javy vrátane reči ako „reflex“. Behaviorizmus si asi raz zaslúži svoj vlastný článok, keďže to bolo prekrútenie racionalizmu; ale toto nie je ten článok.

„Nazývaš to ‚blbosť‘,“ pýtate sa, „ale odkiaľ *vieš*, že si tvoj mozog reprezentuje vizuálne obrazy? Iba preto, že môžeš zavrieť oči a vidieť ich?“

Na túto otázku *bolo* za oných čias kontroverzie veľmi ťažké odpovedať. Ak ste chceli dokázať existenciu myšlienkových obrazov „vedecky“, namiesto iba pomocou introspekcie, museli ste odvodiť existenciu myšlienkových obrazov pomocou pokusov ako napríklad: Ukážte pokusnej osobe dva predmety a opýtajte sa, či jeden z nich možno otočiť tak, aby sa stal tým druhým. Čas reakcie je lineárne

→ http://lesswrong.com/lw/o8/conditional_independence_and_naive_bayes/

úmerný potrebnému uhlu otočenia. Toto je ľahké vysvetliť, ak si naozaj predstavujete ten obrázok a spojite ho otáčate konštantnou rýchlosťou, ale je ťažké vysvetliť, ak si iba kontrolujete proporcie častí obrázku.

Dnes dokážeme naozaj zachytiť malé obrázky v zrakovej kôre. Takže áno, váš mozog naozaj reprezentuje podrobný obrázok toho, čo vidí, alebo si predstavuje. Pozrite si od Stephena Kosslyna *Obraz a mozog: Výsledok debaty o obraznosti*.[172]

Časť dôvodu, prečo majú ľudia problémy so slovami, je neuvedomovanie si, koľko zložitosti sa za slovami skrýva.

Viete si predstaviť „zeleného psa“? Viete si predstaviť „syrové jablko“?

„Jablko“ nie je iba postupnosť dvoch slabík či piatich písmen. Je to tieň. Je to špička tigrieho chvosta.

Slová, alebo skôr pojmy za nimi, sú štetce – môžete ich použiť na kreslenie obrazov vo svojej vlastnej mysli. Doslova na kreslenie, ak použijete pojmy na vytvorenie obrázku vo svojej zrakovej kôre. A použitím spoločných označení môžete siahnuť do mysle niekoho iného a chytiť ich štetec, aby ste nakreslili obrázok v *jeho* mysli – obrázok ako malý zelený pes v jeho zrakovej kôre.

Nemyslite si však, že keď posielate slabiky cez vzduch alebo písmená cez internet, že to tie slabiky alebo písmená kreslia obrázky v zrakovej kôre. To by si vyžadovalo zložité pokyny, ktoré by sa do tejto postupnosti písmen nezmestili. „Jablko“ má 6 bajtov, a nakresliť obrázok jablka od nuly by si vyžadovalo viac údajov.

„Jablko“ je iba nálepka pripojená k skutočnému a bezslovnému pojmu *jablka*, ktorý dokáže nakresliť obrázok vo vašej zrakovej kôre, alebo sa zraziť so „syrom“, alebo rozoznať jablko, keď nejaké uvidíte, alebo ochutnať jeho archetyp v jablkovom koláči, možno aj vyslať pohybové správanie na jedenie jablka...

A nie je to len také jednoduché ako vyvolanie si obrázku z pamäte. Ako by ste si potom dokázali predstavovať kombinácie ako „trojuholníkové žiarovky“ - aplikovať trojuholníkovosť na žiarovku, aj keď ste takú vec nikdy v živote nevideli?

Neurobte chybu, ktorú robili behavioristi. Reč je viac než iba zvuk vo vzduchu. Označenia sú iba ukazovatele - „pozri sa v pamäti do oblasti 1387540“. Skôr či neskôr, keď dostanete ukazovateľ, príde čas dereferencovať ho, a naozaj sa pozrieť v pamäti na oblasť 1387540.

Na čo to slovo ukazuje?

*→

179. Klam otázky s premennou

> Albert: „Vždy, keď počúvam, ako strom padá, vydá zvuk, takže si myslím, že ostatné padajúce stromy tiež vydávajú zvuk. Neverím tomu, že sa svet zmení, keď sa nepozerám.“
>
> Barry: „Počkaj chvíľu. Ak to nikto nepočuje, ako to môže byť zvuk?“

Kým píšem dialóg Alberta a Barryho pri ich debate o tom, či padajúci strom v opustenom lese vydáva zvuk, občas zisťujem, že strácam empatiu so svojimi postavami. Začínam strácať inštinktívne chápanie, prečo by sa *niekto* niekedy takto hádal, hoci som to videl mnohokrát.

V takom prípade si zopakujem: „Buď padajúci strom vydáva zvuk, alebo nie!“, aby som si obnovil svoj vypožičaný pocit rozhorčenia.

(P alebo ~P) nie je vždy spoľahlivá heuristika, ak namiesto P dosadzujete anglické vety. „Táto veta je nepravdivá“ nemôže byť konzistentne vnímaná ako pravdivá ani ako nepravdivá. Alebo stará klasika: „Už si prestal biť svoju ženu?“

---

172 Stephen M. Kosslyn, *Image and Brain: The Resolution of the Imagery Debate* [Obraz a mozog: Výsledok debaty o obraznosti] (Cambridge, MA: MIT Press, 1994).

→ http://lesswrong.com/lw/o9/words_as_mental_paintbrush_handles/

Ak ste matematici a veríte v klasickú (nie intuicionistickú) logiku, máte možnosti, ako naďalej trvať na tom, že platí (P alebo ~P): napríklad povedať, že „táto veta je nepravdivá“ nie je veta.

Ale takéto riešenia sú jemné, čo stačí na ukázanie potreby jemnosti. Nemôžete len tak pri každej príležitosti rúbať okolo seba slovami: „Buď to je, alebo to nie je!“

Takže vydáva padajúci strom zvuk, alebo nie, alebo...?

Určite platí 2 +2 = X, alebo nie? No, možno, *ak* je to *naozaj* to isté X, to isté 2, a to isté + a =. Ak X v niektorých prípadoch vyhodnocujeme ako 5 a v iných ako 4, vaše rozhorčenie môže byť nesprávne adresované.

Aby sme vôbec začali tvrdiť, že (P alebo ~P) by mala byť nevyhnutne pravda, symbol P musí v oboch častiach danej dilemy označovať *presne* tú istú vec. „Buď padajúci strom vydá zvuk alebo nie!“ - ale ak Albert::zvuk nie je to isté ako Barry::zvuk, nie je nič paradoxné na tom, ak strom vydá Albert::zvuk, ale nie Barry::zvuk.

(Zápis :: je niečo, čo som si osvojil za svojich čias C++ na vyhnutie sa konfliktom v priestore názvov. Ak máte dva rôzne balíky, ktoré definujú triedu Zvuk, môžete napísať Balík1::Zvuk, aby ste upresnili, ktorý Zvuk tým myslíte. Tento zápis asi nie je všeobecne známy, čo je škoda, lebo si často želám, aby som ho mohol používať pri písaní.)

Táto premenlivosť môže byť jemná: Albert a Barry môžu pozorne overiť, či je to ten istý strom, v tom istom lese, a tá istá udalosť padania, len aby sa uistili, že naozaj majú podstatnú nezhodu o presne tom istom pojme.

Predstavte si potraviny, v ktorých najčastejšie nakupujete: Sú na ľavej strane cesty, alebo na pravej? No samozrejme neexistuje „*ľavá* strana“ cesty, iba *vaša ľavá* strana, keď idete pozdĺž cesty niektorým smerom. Mnohé zo slov, ktoré používame, sú v skutočnosti funkcie implicitných premenných dodávaných z prostredia.

Je to naozaj veľká námaha vyžadujúca si veľa práce, aby ste zvládli tento druh problému v programe pre umelú inteligenciu na analýzu jazyka – jav, ktorý sa označuje „deixis hovoriaceho“.

„Martin povedal Bobovi, že budova je na jeho ľavej strane.“ Ale „ľavá“ je slovo-funkcia, ktoré sa vyhodnocuje pomocou premennej závisiacej na hovoriacom, ktorú si neviditeľne vyberá z okolitého kontextu. Čia „ľavá“ sa myslí, Bobova alebo Martinova?

Premenné v klame otázky z premennou často nie sú úhľadne označené – nie je to také jednoduché ako: „Počuj, myslíš si, že Z + 2 sa rovná 6?“

Ak zrážka priestorov názvov prinesie dva rôzne pojmy, ktoré vyzerajú ako „ten istý pojem“, pretože majú rovnaké meno – alebo kompresia mapy prinesie dve rôzne udalosti, ktoré vyzerajú ako tá istá udalosť, pretože nemajú samostatné myšlienkové súbory – alebo ak sa tá istá funkcia vyhodnocuje v rôznych kontextoch – potom sa samotná skutočnosť stane premenlivou. Prinajmenšom takto to algoritmu pripadá zvnútra. Oko vašej mysle vidí mapu, nie priamo územie.

Ak máte otázku so skrytou premennou, ktorá sa v rôznych kontextoch vyhodnocuje na rôzne výrazy, *vyzerá to*, akoby samotná skutočnosť bola nestabilná – čo vidí oko vašej mysle sa mení podľa toho, kam sa pozerá.

Toto často pletie doktorandov (a profesorov postmoderny), ktorí zistia, že nejaká veta má viac než jednu interpretáciu; myslia si, že objavili nestabilnú časť skutočnosti.

„Ach, ježkove oči! ‚Slnko obieha okolo Zeme‘ je pravda pre lovca a zberača Hungu, ale pre astronóma Amaru je ‚Slnko obieha okolo Zeme‘ nepravda! Neexistuje žiadna pevne daná pravda!“ Dekonštrukciu tejto prváckej pochabosti ponechávam ako cvičenie pre čitateľa.

A predsa som sa aj ja pôvodne pristihol, ako píšem: „Ak sa X v istých prípadoch rovná 5 a v iných 4, veta ‚2 + 2 = X‘ nemusí mať pevne danú pravdivostnú hodnotu.“ To ale nie je *jedna* veta s *premenlivou* pravdivostnou hodnotou. „2 + 2 = X“ *nemá* pravdivostnú hodnotu. Ešte to nie je *veta*, v tom zmysle ako matematici definujú vety, rovnako ako „2 + 2 =“ nie je veta, ani „Fred preskočil cez“ nie je gramatická veta.

Ale táto chyba ma sklon sa prešmyknúť, ešte aj keď údajne vieme viac, pretože takto to skrátka algoritmu pripadá zvnútra.

*→

180. 37 spôsobov, ako slová môžu byť chybné

Niektorý čitateľ zaručene vyhlási, že lepším nadpisom tejto kapitoly by bolo: „37 spôsobov, ako môžete nerozumne používať slová“ alebo „37 spôsobov, ako suboptimálne používanie kategórií môže mať škodlivé vedľajšie účinky na vaše poznávanie.“

Ale jedna zo základných lekcií tohto gigantického zoznamu je, že povedať „Nie je možné, že by moja voľba X bola ‚chybná‘“ je takmer vždy chybou v praxi, bez ohľadu na teóriu. Vždy môžete urobiť chybu. Ešte aj keď je teoreticky nemožné urobiť chybu, aj tak môžete urobiť chybu. Neexistuje žiadna karta Vyjdi-z-väzenia-zadarmo na všetko, čo urobíte. Taký je život.

Navyše, môžem si definovať slovo „chyba“ tak, aby označovalo čokoľvek, čo sa mi zachce – slovo predsa nemôže byť *chybné*.

Osobne si myslím, že je pomerne oprávnené použiť slovo „chyba“, keď:

1. *Slovo sa v prvom rade nijako nespája so skutočnosťou.* Je Sokrates framster? Áno alebo Nie? (Podobenstvo o dýke.)
2. *Váš argument, keby fungoval, by dokázal prinútiť skutočnosť ísť iným smerom pomocou vybrania odlišnej definície slova.* Sokrates je človek, a ľudia sú podľa definície smrteľní. Keby sme definovali ľudí ako nesmrteľných, žil by Sokrates naveky? (Podobenstvo o bolehlave.)
3. *Pokúšate sa ustanoviť nejaké empirické tvrdenie ako pravdivé „podľa definície“.* Sokrates je človek, a ľudia sú podľa definície smrteľní. Je to teda *logická pravda*, ak empiricky predpovieme, že Sokrates sa zrúti, keď vypije bolehlav? Zdá sa, že sú logicky možné vnútorne nerozporné svety, kde sa Sokrates nezrúti – kde je povedzme vďaka náhodnej biochémii odolný voči bolehlavu. Logické pravdy sú pravdivé vo všetkých možných svetoch, a preto vám nikdy nepovedia, v *ktorom* možnom svete žijete – a čokoľvek, čo dokážete odvodiť „z definície“ je logická pravda. (Podobenstvo o bolehlave.)
4. *Nevedomky ste na niečo prilepili konvenčnú nálepku, a v skutočnosti nepoužívate slovnú definíciu, ktorú ste práve dali.* Dokonale viete, že Bob je „človek“, hoci podľa vašej definície by ste Boba nikdy nemali nazvať „človekom“, dokiaľ najprv nepozorujete jeho smrteľnosť. (Podobenstvo o bolehlave.)
5. *Akt označenia niečoho slovom skrýva spochybniteľné induktívne odvodenia, ktoré robíte.* Ak posledných 11 vytiahnutých predmetov tvaru vajca bolo modrých, a posledných 8 vytiahnutých kociek bolo červených, je vecou indukcie povedať, že toto pravidlo bude platiť aj v budúcnosti. Ale ak nazvete modré vajcia „majcia“ a červené kocky „čocky“, môžete siahnuť do suda, nahmatať vajcovitý tvar a pomyslieť si „Aha, majce.“ (Slová ako skryté odvodenia.)
6. *Pokúšate sa definovať slovo pomocou slov, ktoré potom definujete pomocou ešte abstraktnejších slov, bez ukázania na príklad.* „Čo je to červená?“ „Červená je farba.“ „Čo je to farba?“ „Je to vlastnosť veci.“ „Čo je to vec? Čo je to vlastnosť?“ Nikdy vám nenapadne zastaviť sa a ukázať na značku stop a na jablko. (Extenzie a intenzie.)
7. *Extenzia nezodpovedá intenzii.* Vedome si neuvedomujeme naše stotožnenie červeného svetla na oblohe ako „Mars“, čo sa pravdepodobne stane bez ohľadu na váš pokus definovať „Mars“ ako „boh vojny“. (Extenzie a intenzie.)
8. *Vaša slovná definícia nezachytáva viac než maličký zlomok spoločných vlastností danej kategórie, ale snažíte sa argumentovať, akoby to tak bolo.* Keď filozofi Platónovej Akadémie tvrdili, že najlepšia definícia človeka je „dvojnožec bez peria“, Diogenes Cynik údajne ukázal ošklbané kura a vyhlásil: „Toto je Platónov človek.“ Platónovci pohotovo zmenili svoju definíciu na „dvojnožec bez peria so širokými nechtami.“ (Zhluky podobnosti.)

→ http://lesswrong.com/lw/oc/variable_question_fallacies/

9. *Pokúšate sa brať členstvo v kategórii ako všetko alebo nič, ignorujúc existenciu viac a menej typických podzhlukov.* Kačice a tučniaky sú menej typické vtáky ako červienky a holuby. Zaujímavé je, že pokus medzi skupinami ukázal, že pokusné osoby si myslia, že choroba sa na ostrove skôr rozšíri z červienok na kačice než z kačíc na červienky. (Typickosť a asymetrická podobnosť.)

10. *Slovná definícia funguje v praxi dosť dobre na to, aby ukázala na zamýšľaný zhluk podobných vecí, ale riešite výnimky.* Nie každý človek má desať prstov, nie každý nosí šaty alebo používa jazyk; ale keď sa pozriete na empirický zhluk vecí, ktoré majú spoločné tieto vlastnosti, dostanete dosť informácií na to, aby vás občasný deväťprstý človek nepoplietol. (Zhluková štruktúra priestoru vecí.)

11. *Pýtate sa, či niečo „je" alebo „nie je" členom kategórie, ale neviete pomenovať otázku, na ktorú naozaj chcete odpoveď.* Čo je to „muž"? Je bábätko Barney „muž"? „Správna" odpoveď môže výrazne závisieť od toho, či *naozaj* hľadáte odpoveď na otázku: „Bolo by dobré ponúknuť Barneymu bolehlav?" alebo „Bol by Barney dobrým manželom?" (Skryté otázky.)

12. *Považujete intuitívne vnímané hierarchické kategórie za jediný správny spôsob analýzy sveta a neuvedomujete si, že sú možné aj iné druhy štatistickej inferencie, aj keď ich váš mozog nepoužíva. Človeku* je omnoho ľahšie všimnúť si, či je objekt „majce" alebo „čocka" než všimnúť si, že červené predmety nikdy nesvetielkujú v tme, ale červené chlpaté predmety majú všetky zvyšné vlastnosti rovnaké ako majcia. Iné štatistické algoritmy fungujú inak. (Neurónové kategórie.)

13. *Hovoríte o kategóriách, akoby to bola manna spadnutá z Platónovskej Ríše namiesto odvodení implementovaných v skutočnom mozgu.* Starovekí filozofi hovorili „Sokrates je človek", nie „môj mozog na základe vnemov klasifikuje Sokrata ako zodpovedajúceho pojmu ‚človek'." (Ako sa algoritmus cíti zvnútra.)

14. *Argumentujete o členstve v kategórii ešte aj pod oddelení všetkých zvyšných otázok, ktoré by mohli závisieť na úsudku podľa kategórie.* Keď už ste videli, že nejaký predmet je modrý, v tvare vajca, chlpatý, pružný, nepriehľadný, svetielkujúci a obsahuje paládium, na čo *ešte* sa pýtate slovami: „Je to majce?" Ale ak kategorizujúca neurónová sieť vášho mozgu obsahuje (metaforický) centrálny uzol zodpovedajúci odvodeniam z majcovitosti, stále sa *zdá*, akoby nám tu zostala nezodpovedaná otázka. (Ako sa algoritmus cíti zvnútra.)

15. *Dovolíte, aby sa diskusia otočila na definície, hoci to nie je to, o čom ste sa bavili pôvodne.* Keby ste sa pred začiatkom diskusie o tom, či strom padajúci v opustenom lese vydáva zvuk, opýtali budúcich hádajúcich sa, či si myslia, že by sa „zvuk" mal definovať ako „akustické vibrácie" alebo „sluchovné vnemy", pravdepodobne by vám povedali, nech si hodíte mincou. Až po začiatku hádky sa definícia slova stane politicky nabitou. (Debaty o definíciách.)

16. *Myslíte si, že slovo má význam, ako vlastnosť samotného slova; namiesto toho, že je to nálepka, ktorú si váš mozog spája s konkrétnym pojmom.* Ked niekto skríkne: „Jaj! Tiger!", evolúcia neuprednostní organizmus, ktorý si pomyslí: „Hm... práve som počul slabiky ‚ti' a ‚ger', ktoré moju súkmeňovci spájajú so svojimi vnútornými analógmi môjho pojmu *tigra* a ktoré *aíííí* CHRUM CHRUM GLG." Mozog teda ide skratkou a zdá sa, že význam tigrovitosti je vlastnosťou samotnej nálepky. Ľudia sa potom hádajú o *správnom význame* nálepky ako „zvuk". (Precíť význam.)

17. *Hádate sa o význame slova, hoci obe strany dokonale rozumejú, čo sa druhá strana snaží povedať.* Ľudská schopnosť spájať si nálepky s pojmami je nástroj na komunikáciu. Keď ľudia *chcú* komunikovať, je ťažké nás zastaviť; ak nemáme spoločný jazyk, budeme si kresliť obrázky do piesku. Keď každý z vás rozumie, čo ten druhý myslel, je *vybavené*. (Argument bežným používaním.)

18. *Uprostred empirickej alebo morálnej debaty vytiahnete slovník.* Redaktori slovníkov sú historikmi používania, nie zákonodarcami jazyka. Ak má bežná definícia problém – ak je „Mars" definovaný ako boh vojny, alebo „delfín" ako druh ryby, alebo „černoch" ako samostatná

kategória mimo človeka, slovník bude odrážať túto štandrdnú chybu. (Argument bežným používaním.)

19. *Uprostred ľubovoľnej debaty vytiahnete slovník.* Vážne, na základe čoho si myslíte, že redaktori slovníka sú autoritami na otázku, či „ateizmus“ je „náboženstvo“, alebo na čokoľvek iné? Ak vám ide o nejakú vecnú záležitosť, naozaj si myslíte, že redaktori slovníkov majú prístup ku konečnej múdrosti, ktorá túto debatu vyrieši? (Argument bežným používaním.)

20. *Bezdôvodne vzdorujete bežnému používaniu, čím druhým svojvoľne komplikujete chápanie.* Rýchle plutónium s rohlíkom bez rúčky. (Argument bežným používaním.)

21. *Používate zložité premenovania, aby ste vytvorili ilúziu odvodenia.* Je „človek“ definovaný ako „smrteľný dvojnožec bez peria“? Potom napíšte: „Každý [smrteľný dvojnožec bez peria] je smrteľný. Sokrates je [smrteľný dvojnožec bez peria]; preto Sokrates je smrteľný.“ Vyzerá to tak menej pôsobivo, však? (Prázdne nálepky.)

22. *Dostanete sa do hádky, ktorej by ste sa mohli vyhnúť, keby ste skrátka nepoužili dané slovo.* Keby mali Albert a Barry zakázané použiť slovo „zvuk“, potom by Albert musel povedať: „Strom padajúci v opustenom lese vytvára akustické vibrácie“ a Barry by povedal: „Strom padajúci v opustenom lese nevyvoláva sluchové vnemy.“ Keď slovo predstavuje problém, najjednoduchším riešením je odstrániť slovo a jeho synonymá. (Hrajte sa so slovami na tabu.)

23. *Existencia pekného malého slovíčka vám bráni vidieť podrobnosti veci, o ktorej sa pokúšate rozmýšľať.* Čo sa naozaj odohráva v školách, keď to prestanete nazývať „vzdelávanie“? Čo je titul, keď ho prestanete nazývať „titul“? Ak na minci padne „hlava“, aká je jej radiálna orientácia? Čo je to „pravda“, ak nemôžete povedať „presné“ ani „správne“ ani „predstavuje“ ani „odráža“ ani „sémantické“ ani „veriť“ ani „vedomosť“ ani „mapa“ ani „skutočné“ ani žiadne ďalšie jednoduché slovo? (Nahraďte symbol podstatou.)

24. *Máte iba jedno slovo, ale v skutočnosti sú to dve alebo viac rôznych vecí, takže všetky fakty o nich sa dostanú do jedného nerozlíšeného myšlienkového vedierka.* Je súčasťou bežnej detektívovej práce všimúť si, že Carol bola minulú noc oblečená v červenom, alebo že má čierne vlasy; a je súčasťou bežnej detektívovej práce zamyslieť sa, či si Carol nefarbí vlasy. Ale treba jemnejšieho detektíva, aby sa zamyslel, či nie sú dve Carol, takže Carol, ktorá chodí v červenom nemusí byť tá istá ako Carol, ktorá mala čierne vlasy. (Chyby kompresie.)

25. *Vidíte vzory tam, kde žiadne neexistujú, ťažíte zo svojich definícií ďalšie vlastnosti, aj keď v tomto smere nie je žiadna podobnosť.* V Japonsku si myslia, že ľudia s krvnou skupinou A sú úprimní a tvoriví, s krvnou skupinou B sú divokí a veselí, s krvnou skupinou 0 sú príjemní a spoločenskí, a s krvnou skupinou AB sú pokojní a vyrovnaní. (Kategorizovanie má dôsledky.)

26. *Pokúšate sa prepašovať konotáciu nejakého slova argumentovaním definíciou, ktorá túto konotáciu neobsahuje.* „Wiggin“ je v slovníku definovaný ako osoba so zelenými očami a čiernymi vlasmi. Slovo „wiggin“ má však konotáciu niekoho, kto pácha zločiny a vystreľuje malé veveričky, ale táto časť v slovníku nie je. Takže ukážete na niekoho a poviete: „Zelené oči? Čierne vlasy? Hovorím ti, že je to wiggin! Sleduj ako bude kradnúť strieborné príbory.“ (Prepašovanie konotácií.)

27. *Tvrdíte, že „X je podľa definície Y!“ V takom prípade sa takmer určite pokúšate prepašovať konotáciu Y, ktorá nebola v pôvodnej definícii.* Definujete si „človeka“ ako „dvojnožca bez peria“ a ukážete na Sokrata so slovami: „Dve nohy, žiadne perie – musí to byť človek!“ Ale o čo vám *naozaj* ide, je niečo iné, napríklad smrteľnosť. Keby bola debata o počte Sokratových nôh, tam by ten druhý odpovedal iba: „To máš odkiaľ, že Sokrates má dve nohy? O tom sa predsa od začiatku hádame!“ (Argumentovanie „podľa definície“.)

28. *Tvrdíte „P sú podľa definície Q!“* Ak vidíte Sokrata na poli s nejakými biológmi, ako zbiera rastliny, ktoré môžu poskytovať odolnosť voči bolehlavu, nemá zmysel argumentovať: „Ľudia sú podľa definície smrteľní!“ Najčastejšie cítite potrebu pritiahnuť zverák trvaním na tom, že niečo je pravda „podľa definície“, keď sú tu iné informácie, ktoré toto štandardné odvodenie spochybňujú. (Argumentovanie „podľa definície“.)

29. *Pokúšate sa ustanoviť členstvo v empirickom zhluku „podľa definície“*. Nebudete cítiť potrebu hovoriť: „Hinduizmus je *z definície* náboženstvo!“ pretože je samozrejmé, že hinduizmus je náboženstvo. Nie je to iba náboženstvo „podľa definície“, ale je to *naozajstné* náboženstvo. Ateizmus sa nepodobá na centrálnych členov zhluku „náboženstvo“, takže keby to nebolo tak, že ateizmus je náboženstvo *podľa definície*, človek by si mohol myslieť, že ateizmus *nie je* náboženstvo. Preto potrebujete drviť odporcov ukazovaním na to, že „ateizmus je náboženstvo“ je pravda *podľa definície*, pretože to nie je pravda podľa ničoho iného. (Argumentovanie „podľa definície“.)

30. *Vaša definícia kreslí hranicu okolo vecí, ktoré v skutočnosti nepatria dokopy.* Môžete tvrdiť, ak chcete, že definujete slovo „ryba“ tak, že odkazuje na lososa, gupky, žraloky, delfíny a pstruhy, ale nie na medúzy a riasy. Môžete tvrdiť, ak chcete, že toto je jednoducho zoznam, a zoznam skrátka nemôže byť „nesprávny“. Alebo môžete prestať hrať priblblé hry a pripustiť, že ste urobili chybu, a že delfíny do zoznamu rýb nepatria. (Kde nakresliť hranicu?)

31. *Môžete použiť krátke slovo na niečo, čo nepotrebujete opisovať často, alebo dlhé slovo na niečo, čo potrebujete opisovať často. Toho výsledkom môže byť neefektívne myslenie, alebo zlé použitie Occamovej britvy, ak si vaša myseľ myslím, že krátke vety znejú „jednoduchšie“*. Čo znie uveriteľnejšie: „Boh urobil zázrak“ alebo „Nadprirodzená bytosť tvoriaca vesmír dočasne pozastavila fyzikálne zákony“? (Entropia a krátke kódy.)

32. *Kreslíte hranicu okolo objemu priestoru, kde nie je vyššia než zvyčajná hustota, čo znamená, že príslušné slovo nezodpovedá žiadnemu vykonateľnému bayesovskému odvodeniu.* Keďže zelenookí ľudia nemajú väčšiu pravdepodobnosť mať čierne vlasy, ani naopak, a nemajú žiadne ďalšie spoločné vlastnosti, načo je vôbec slovo „wiggin“? (Vzájomná informácia a hustota v priestore vecí.)

33. *Bezdôvodne kreslíte komplikovanú hranicu.* Akt definovania slova, ktoré zodpovedá všetkým ľuďom okrem černochov vyzerá trochu podozrivo. Ak nepredložíte dôvody na nakreslenie tejto konkrétnej hranice, pokúšať sa vytvoriť „svojvoľné“ slovo v tejto oblasti je ako keď detektív povie: „Nuž, nemám najmenšiu štipku podpory jedným ani druhým smerom, kto mohol zavraždiť tieto siroty... ale zamysleli sme sa nad tým, či John Q. Wiffleheim nie je podozrivý?“ (Superexponenciálny priestor pojmov a jednoduché slová.)

34. *Používate kategorizáciu na to, aby ste vytvorili odvodenia vlastností, ktoré nemajú správnu empirickú štruktúru, konkrétne, podmienenú nezávislosť pri danej vedomosti o triede, ktorá sa dá napodobniť naivným Bayesom.* Nie, toto nezhrniem jednou vetou. Prečítajte si danú kapitolu. (Podmienená nezávislosť a naivný Bayes.)

35. *Myslíte si, že slová sú ako maličké symboly LISPu vo vaše mysli, namiesto toho, že slová sú nálepky, ktoré fungujú ako rúčky na ovládanie zložitých myšlienkových štetcov, ktoré dokážu kresliť podrobné obrázky vo vašom zmyslovom pracovisku.* Predstavte si „trojuholníkovú žiarovku“. Čo ste videli? (Slová ako rúčky myšlienkových štetcov.)

36. *Používate slovo, ktoré má na rôznych miestach rôzne významy, akoby znamenalo vždy tú istú vec, čím možno vytvárate ilúziu niečoho premenlivého a pohyblivého.* „Martin povedal Bobovi, že budova je na jeho ľavej strane.“ Ale „ľavá“ je slovo-funkcia, ktoré sa vyhodnocuje podľa premennej závislej od hovoriaceho, vybranej z okolitého kontextu. Čia „ľavá“ sa myslela, Bobova alebo Martinova? (Klam otázky s premennou.)

37. *Myslíte si, že definície nemôžu byť „nesprávne“ alebo že „môžem definovať slovo, ako sa mi zachce!“* Tento postoj vás učí rozhorčene brániť svoje minulé činy, namiesto venovania pozornosti ich dôsledkom, alebo priznania si svojich chýb. (37 spôsobov, ako subptimálne použitie kategórií môže mať škodlivé vedľajšie účinky na vaše poznávanie.)

Všetko, čo robíte v mysli, má nejaký účinok a váš mozog nevedome uháňa vpred bez vášho dozoru.

Povedať „Slová sú svojvoľne dané; môžem si definovať slovo, ako sa mi zachce“ dáva asi rovnako veľa zmyslu ako šoférovať auto na poľadovici s plynovým pedálom na podlahe a hovoriť: „Ako pozerám

na tento volant, nevidím prečo je táto konkrétna poloha špeciálna – môžem si teda otočiť volantom, ako sa mi zachce.“

Ak sa pokúšate niekam ísť, alebo sa čo len pokúšate *prežiť*, radšej by ste mali začať venovať pozornosť trom či šiestim tuctom kritérií optimálnosti, ktoré určujú, ako používať slová, definície, kategórie, triedy, hranice, nálepky a pojmy.

*→

Medzihra: Intuitívne vysvetlenie Bayesovej vety

Redakčná poznámka: Toto je skrátená verzia pôvodného článku, ktorý obsahoval mnoho interaktívnych prvkov.

Vaši priatelia a kolegovia hovoria o niečom, čo sa volá „Bayesova veta“ alebo „Bayesovo pravidlo“, alebo o niečom, čo sa volá bayesovské usudzovanie. Znejú tým naozaj nadšení, takže si pogooglite a nájdete webovú stránku o Bayesovej vete a...

Je to táto rovnica. To je všetko. Iba jedna rovnica. Stránka vám dá jej definíciu, ale nehovorí o tom, čo je to, prečo je užitočná, alebo prečo sa o ňu vaši priatelia asi zaujímajú. Vyzerá to ako náhodná vec zo štatistiky.

Prečo jeden matematický pojem vyvoláva takéto čudné nadšenie v jeho študentoch? Čo je tá takzvaná bayesovská revolúcia, ktorá sa teraz šíri vedou a tvrdí, že dokonca samotnú experimentálnu metódu zahŕňa v sebe ako špeciálny prípad? Čo je to tajomstvo, ktoré poznajú Bayesovi prívrženci? Čo je to svetlo, ktoré uvideli?

Čoskoro budete vedieť. Čoskoro budete jedným z nás.

Hoci na internete existuje pár vysvetlení Bayesovej vety, mám skúsenosť, keď sa pokúšam ľudí uviesť do bayesovského uvažovania, že existujúce internetové vysvetlenia sú príliš abstraktné. Bayesovské uvažovanie je veľmi *kontraintuitívne*. Ľudia nepoužívajú bayesovské uvažovanie intuitívne, pripadá im veľmi náročné naučiť sa bayesovské uvažovanie počas výcviku, a rýchlo zabúdajú na bayesovské metódy, keď výcvik skončí. Toto platí rovnako pre začínajúcich študentov aj pre vysoko trénovaných profesionálov z praxe. Bayesovské uvažovanie je zrejme jedna z tých vecí ako kvantová mechanika alebo Wasonov test výberu, ktoré sú pre ľudí ťažké pochopiť našimi prirodzenými myšlienkovými schopnosťami.

Alebo sa tak tvrdí. Tu nájdete pokus o poskytnutie *intuitívneho* bayesovského uvažovania – neznesiteľne jemný úvod, ktorý zahŕňa všetky ľudské spôsoby chápania čísel, od prirodzených frekvencií po priestorové vizualizácie. Zámerom je sprostredkovať nie abstraktné pravidlá na narábanie s číslami, ale čo tieto čísla znamenajú, a prečo sú pravidlá také, aké sú (a nemôžu byť iné). Keď dočítate túto stránku, budete sa vám o bayesovských problémoch snívať.

Tak začnime.

* * *

Tu je príbeh o situácii, s ktorou sa lekári často stretávajú:

> 1 % žien vo veku štyridsať rokov, ktoré sa zúčastňujú preventívneho vyšetrenia, má rakovinu prsníka. 80 % žien s rakovinou prsníka dostane pozitívne mamogramy. 9,6 % žien bez rakoviny prsníka tiež dostane pozitívne mamogramy. Žena z tejto vekovej skupiny mala pozitívny mamogram pri preventívnom vyšetrení. Aká je pravdepodobnosť, že naozaj má rakovinu prsníka?

Čo si myslíte, že je odpoveď? Ak ste sa s problémom tohto druhu ešte nestretli, prosím pokúšajte sa chvíľu dôjsť na vlastnú odpoveď, než budete pokračovať.

* * *

Ďalej si predstavte, že vám poviem, že väčšina lekárov pri tomto probléme dostane zlú odpoveď – zvyčajne iba asi 15 % doktorov odpovie správne. („Naozaj? 15 %? To je skutočné číslo, alebo je to

→ http://lesswrong.com/lw/od/37_ways_that_words_can_be_wrong/

legenda založená na internetovej ankete?“ Je to skutočné číslo. Viď Casscells, Schoenberger, a Grayboys 1978;[173] Eddy 1982;[174] Gigerenzer a Hoffrage 1995;[175] a mnohé ďalšie štúdie. Je to prekvapujúci výsledok, ktorý sa ľahko replikuje, takže ho mnohokrát replikovali.)

Pri horeuvedenom príbehu väčšina lekárov odhaduje pravdepodobnosť medzi 70 % a 80 %, čo je veľmi nesprávne.

Tu je alternatívna verzia daného problému, pri ktorej sa doktorom darí o čosi lepšie:

> 10 z 1000 žien vo veku štyridsať rokov, ktoré sa zúčastnili na preventívnom vyšetrení, má rakovinu prsníka. 800 z 1000 žien s rakovinou prsníka dostane pozitívne mamogramy. 96 z 1000 žien bez rakoviny prsníka tiež dostane pozitívne mamogramy. Ak 1000 žien z tejto vekovej skupiny absolvuje preventívne vyšetrenie, aké asi percento žien s pozitívnymi mamogrammi bude mať naozaj rakovinu prsníka?

A na záver, tu je problém, na ktorom sa lekárom darilo najlepšie zo všetkých, kde 46 % - takmer polovica – došla k správnej odpovedi:

> 100 z 10 000 žien vo veku štyridsať rokov, ktoré sa zúčastnili na preventívnej prehliadke, má rakovinu prsníka. 80 z každých 100 žien s rakovinou prsníka dostane pozitívny mamogram. 950 z 990 žien bez rakoviny prsníka dostane pozitívny mamogram. Ak 10 000 žien v tejto vekovej skupine pôjde na preventívnu prehliadku, aké percento žien s pozitívnymi mamogrammi bude mať naozaj rakovinu prsníka?

* * *

Správna odpoveď je 7,8 %, vypočítaná takto: Spomedzi 10 000 žien 100 má rakovinu prsníka, z týchto 100 má 80 pozitívny mamogram. Z tých istých 10 000 žien 9 900 nemá rakovinu prsníka, a z týchto 9 900 žien má 950 tiež pozitívny mamogram. To znamená, že celkový počet žien s pozitívnym mamogramom je 950 + 80, čiže 1 030. Z týchto 1 030 žien s pozitívnym mamogramom má 80 rakovinu. Vyjadrené ako pomer je to 80 / 1030 alebo 0,07767 alebo 7,8 %.

Povedané inými slovami, pred vyšetrením na mamografe môžeme týchto 10 000 žien rozdeliť na dve skupiny:

- Skupina 1: 100 žien *s* rakovinou prsníka.
- Skupina 2: 9 900 žien *bez* rakoviny prsníka.

Tieto dve skupiny dávajú spolu 10 000 pacientiek, čím sa potvrdzuje, že sa v matematike žiadne nestratili. Po mamografii môžeme tieto ženy rozdeliť do štyroch skupín:

- Skupina A: 80 žien *s* rakovinou prsníka a s *pozitívnym* mamogramom.
- Skupina B: 20 žien *s* rakovinou prsníka a s *negatívnym* mamogramom.
- Skupina C: 950 žien *bez* rakoviny prsníka a s *pozitívnym* mamogramom.
- Skupina D: 8950 žien *bez* rakoviny prsníka a s *negatívnym* mamogramom.

Súčet skupín A a B, skupín s rakovinou prsníka, zodpovedá skupine 1; a súčet skupín C a D, skupín bez rakoviny prsníka, zodpovedá skupine 2. Ak poskytnete mamografiu 10 000 pacientkam, potom sa medzi 1030 pacientkami s pozitívnym mamogramom nachádza 80 pacientok s pozitívnym mamogramom a rakovinou. Toto je správna odpoveď, ktorú by lekár mal dať pacientke s pozitívnym mamogramom, ak sa opýta, aká je šanca, že má rakovinu prsníka; ak sa toto opýta trinásť pacientok, približne jedna z týchto trinástich bude mať rakovinu.

* * *

Najčastejšou chybou je ignorovať pôvodný podiel žien s rakovinou prsníka, a podiel žien bez rakoviny prsníka, ktoré dostane falošné pozitívne výsledky, a sústrediť sa iba na podiel žien

173 Ward Casscells, Arno Schoenberger, and Thomas Graboys, „Interpretation by Physicians of Clinical Laboratory Results,“ *New England Journal of Medicine* 299 (1978): 999–1001.

174 David M. Eddy, „Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities,“ in *Judgement Under Uncertainty: Heuristics and Biases*, ed. Daniel Kahneman, Paul Slovic, and Amos Tversky (Cambridge University Press, 1982).

175 Gerd Gigerenzer and Ulrich Hoffrage, „How to Improve Bayesian Reasoning without Instruction: Frequency Formats,“ *Psychological Review* 102 (1995): 684–704.

s rakovinou prsníka, ktoré dostanú pozitívne výsledky. Napríklad väčšina lekárov v týchto štúdiách si myslela, že ak okolo 80 % žien s rakovinou prsníka dostane pozitívny mamogram, potom pravdepodobnosť, že žena s pozitívnym mamogramom má rakovinu prsníka musí byť okolo 80 %.

Zistenie výslednej odpovede si vždy vyžaduje *všetky tri* kusy informácie – percento žien s rakovinou prsníka, percento žien bez rakoviny prsníka, ktoré dostané falošné pozitívne výsledky, a percento žien s rakovinou prsníka, ktoré dostanú (správne) pozitívne výsledky.

Pôvodný pomer pacientok s rakovinou prsníka sa označuje ako *pôvodná pravdepodobnosť*. Šanca, že pacientka s rakovinou prsníka dostane pozitívny mamogram, a šanca, že pacientka bez rakoviny prsníka dostane pozitívny mamogram, sa nazývajú *podmienené pravdepodobnosti*. Táto úvodná informácia sa dokopy označuje ako *pôvodné údaje*. Záverečná odpoveď – odhad pravdepodobnosti, že pacientka má rakovinu prsníka, ak vieme, že dostala pozitívny mamogram – sa nazýva *upravená pravdepodobnosť* alebo *výsledná pravdepodobnosť*. Práve sme si ukázali, že výsledná pravdepodobnosť čiastočne závisí od pôvodnej pravdepodobnosti.

Aby ste videli, ako výsledná odpoveď závisí od pôvodného percenta žien s rakovinou prsníka, predstavte si alternatívny vesmír, v ktorom iba jedna žena z milióna má rakovinu prsníka. Aj keby mamografia v tomto svete odhalila rakovinu prsníka v 8 z 10 prípadov, zatiaľ čo by vrátila falošný pozitívny výsledok pre ženu bez rakoviny prsníka iba v 1 z 10 prípadov, stále by bolo stotisíc falošných pozitívnych prípadov na každý skutočný prípad rakoviny. Pôvodná pravdepodobnosť, že žena má rakovinu je taká extrémne nízka, že hoci pozitívny výsledok mamogramu *zvyšuje* odhadovanú pravdepodobnosť, táto pravdepodobnosť sa nezvyšuje na istotu, ba ani na „viditeľnú šancu“; táto pravdepodobnosť sa zvyšuje z 1 : 1 000 000 na 1 : 100 000.

Toto znázorňuje, že výsledok mamografie *nenahrádza* vašu pôvodnú informáciu o pravdepodobnosti, že pacientka má rakovinu; mamografia *posúva* očakávanú pravdepodobnosť v smere výsledku. Pozitívny výsledok posúva pôvodnú pravdepodobnosť smerom nahor; negatívny výsledok posúva pravdepodobnosť nadol. Napríklad v pôvodnom probléme, kde 1 % žien má rakovinu, 80 % žien s rakovinou dostane pozitívny mamogram, a 9,6 % žien bez rakoviny dostane pozitívny mamogram, pozitívny výsledok mamografie *posúva* pravdepodobnosť 1 % nahor na 7,8 %.

Väčšina ľudí pri prvom stretnutí s týmto typom problému vykoná myšlienkovú operáciu *nahradenia* pôvodnej pravdepodobnosti 1 % pravdepodobnosťou 80 %, že žena s rakovinou dostane pozitívny mamogram. Môže to vyzerať ako dobrý nápad, ale jednoducho to nefunguje. „Pravdepodobnosť, že žena s pozitívnym mamogramom má rakovinu prsníka“ nie je vôbec to isté ako „pravdepodobnosť, že žena s rakovinou prsníka má pozitívny mamogram“; to je ako jablká a hrušky.

* * *

Otázka: Prečo bayesovec prešiel cez cestu?

Odpoveď: Na zodpovedanie tejto otázky potrebujete viac informácií.

* * *

Predstavme si, že sud obsahuje mnoho malých plastových vajec. Niektoré vajcia sú namaľované na červeno a niektoré na modro. 40 % vajec v sude obsahuje perly, 60 % neobsahuje nič. 30 % vajec obsahujúcich perly je namaľovaných na modro, aj 10 % vajec neobsahujúcich nič je namaľovaných na modro. Aká je pravdepodobnosť, že modré vajce obsahuje perlu? V tomto príklade sú čísla dosť jednoduché, takže to môžete spočítať v hlave, a odporúčam, aby ste to skúsili.

Kompaktnejší spôsob ako popísať tento problém:

p(perla) = 40 %

p(modré | perla) = 30 %

p(modré | ~perla) = 10 %

p(perla | modré) = ?

Symbol „~“ je skratka pre „nie“, čiže ~perla znamená „nie perla“.

Zápis „modré | perla“ je skratka pre „modré, ak je perla“ alebo „pravdepodobnosť, že vajce je namaľované na modro, ak je dané, že vajce obsahuje perlu“. Položka na pravej strane je to, čo *už viete* alebo *predpokladáte*, a položka na ľavej strane je *dôsledok* alebo *záver*. Ak máme p(modré | perla) = 30 %, a ak *už vieme*, že nejaké vajce obsahuje perlu, potom môžeme *uzavrieť*, že je šanca 30 %, že toto vajce je namaľované na modro. Preto ten záverečný údaj, ktorý hľadáme - „šanca, že modré vajce obsahuje perlu“ alebo „pravdepodobnosť, že vajce obsahuje perlu, ak vieme, že vajce je namaľované na modro“ - píšeme p(perla | modré).

40 % vajec obsahuje perly, a 60 % vajec neobsahuje nič. 30 % vajec obsahujúcich perly je namaľovaných na modro, takže celkovo 12 % vajec obsahuje perlu a je namaľovaných na modro. 10 % vajec neobsahujúcich nič je namaľovaných na modro, takže celkovo 6 % vajec neobsahuje nič a je namaľovaných na modro. Celkovo 18 % vajec je namaľovaných na modro, čiže šanca, že modré vajce obsahuje perlu je 12 / 18 alebo 2 / 3 alebo zhruba 67 %.

Tak ako predtým, vidíme nevyhnutnosť všetkých troch kúskov informácie pri zvažovaní extrémnych prípadov. V (obrovskom) sude, kde iba jedno vajce z tisíca obsahuje perlu, informácia, že vajce je namaľované na modro, posúva pravdepodobnosť z 0,1 % na 0,3 % (namiesto posunutia pravdepodobnosti zo 40 % na 67 %). Podobne, ak 999 vajec z 1000 obsahuje perly, informácia, že vajce je namaľované na modro, posúva pravdepodobnosť z 99,9 % na 99,966 %; pravdepodobnosť, že vajce *neobsahuje* perlu ide z 1 / 1000 na zhruba 1 / 3000.

Pri probléme s perlou a vajcom, väčšina opýtaných, ktorí neboli oboznámení s bayesovským uvažovaním, by pravdepodobne odpovedala, že pravdepodobnosť, že modré vajce obsahuje perlu je 30 %, alebo možno 20 % (šanca 30 % skutočného pozitívneho výsledku mínus šanca 10 % falošného pozitívneho výsledku). Hoci táto myšlienková operácia vyzerá v danú chvíľu ako dobrý nápad, nedáva zmysel z hľadiska položenej otázky. Je to ako pokus, v ktorom sa pýtate druhákov: „Ak osemnásť ľudí nastúpi do autobusu, a potom ďalších sedem ľudí nastúpi do autobusu, koľko rokov má šofér autobusu?“ Mnoho druhákov odpovie: „Dvadsať päť.“ Rozumejú, že sa od nich žiada, aby vykonali nejakú myšlienkovú procedúru, ale ešte si celkom nespojili túto procedúru so skutočnosťou. Podobne, aby sme našli pravdepodobnosť, že žena s pozitívnym mamogramom má rakovinu prsníka, nedáva žiaden zmysel *nahradiť* pôvodnú pravdepodobnosť, že žena má rakovinu, pravdepodobnosťou, že žena s rakovinou prsníka dostane pozitívny mamogram. Ani nemôžete odčítať pravdepodobnosť falošného pozitívneho výsledku od pravdepodobnosti skutočného pozitívneho výsledku. Tieto operácie sú divoko nepodstatné, tak ako sčítavanie ľudí v autobuse pri hľadaní veku šoféra autobusu.

* * *

Štúdia od Gigerenzera a Hoffrageho v 1995 ukázala, že niektoré spôsoby formulovania problémov v príbehoch viac vyvolávajú správne bayesovské uvažovanie.[176] *Najmenej* vyvolávajúce formulácie používali pravdepodobnosti. O čosi viac vyvolávajúce formulácie používali namiesto pravdepodobností frekvencie; problém zostal rovnaký, iba namiesto povedania, že 1 % žien má rakovinu prsníka, sa povedalo, že 1 žena zo 100 má rakovinu prsníka, a že 80 žien zo 100 s rakovinou prsníka by dostali pozitívny mamogram, a tak ďalej. Prečo pri tomto probléme väčší podiel pokusných osôb prejavil bayesovské uvažovanie? Asi preto, lebo povedať „1 žena zo 100“ vás podnecuje predstaviť si X žien s rakovinou, čo vedie k predstave X žien s rakovinou a pozitívnym mamogramom, atď.

Najúčinnejšia zatiaľ nájdená prezentácia je niečo, čo sa nazýva *prirodzené frekvencie* – povedať, že 40 vajec zo 100 obsahuje perly, 12 zo 40 vajec obsahujúcich perly je namaľovaných na modro, a 6 zo 60 vajec neobsahujúcich perly je namaľovaných na modro. Prezentácia *prirodzených frekvencií* je taká, kde sú informácie o pôvodnej pravdepodobnosti zahrnuté v prezentovaní podmienených pravdepodobností. Keby ste sa o podmienených pravdepodobnostiach vajec učili pomocou prirodzených experimentov, tak by ste – v procese rozbíjania sto vajec – rozbili zhruba 40 vajec obsahujúcich perly, z ktorých 12 vajec by bolo namaľovaných na modro, zatiaľ čo by ste rozbili 60 vajec neobsahujúcich nič, z ktorých zhruba 6 by

176 Tamtiež.

bolo namaľovaných na modro. V procese učenia sa podmienených pravdepodobností by ste videli príklady modrých vajec obsahujúcich perly asi dvakrát častejšie než by ste videli príklady modrých vajec neobsahujúcich nič.

Nanešťastie, hoci sú prirodzené frekvencie krokom správnym smerom, pravdepodobne to nebude stačiť. Keď sa problémy predložia v prirodzených frekvenciách, podiel ľudí, ktorí použijú bayesovské uvažovanie, stúpne asi na polovicu. Je to veľké zlepšenie, ale nie dosť veľké, keď hovoríme o skutočných lekároch a skutočných pacientoch.

* * *

Otázka: Ako nájdem pôvodné údaje pre problém?

Odpoveď: Mnohé často používané pôvodné údaje sú uvedené v *chemických a fyzikálnych tabuľkách.*

Otázka: Odkiaľ vlastne *pôvodne* pochádzajú pôvodné údaje?

Odpoveď: Túto otázku sa nikdy nepýtaj.

Otázka: Ehm. Tak odkiaľ majú vedci svoje pôvodné údaje?

Odpoveď: Pôvodné údaje pre vedecké problémy sa určujú každoročným hlasovaním Americkej akadémie vied. V posledných rokoch boli tieto voľby nejednotné a kontroverzné, ovplyvnené nevraživosťou, rozdelením na frakcie, a niekoľkými úkladnými vraždami. Môže to byť odraz vnútorného konfliktu v Bayesovskej rade, alebo dôsledok toho, že debatujúci majú priveľa voľného času. Nikto si nie je istý.

Otázka: Chápem. A odkiaľ majú všetci ostatní svoje pôvodné údaje?

Odpoveď: Pôvodné údaje si sťahujú cez torrenty.

Otázka: Čo ak pôvodné údaje nie sú dostupné na torrentoch?

Odpoveď: V zadnej uličke čínskej štvrte v San Franciscu je malý prepchaný obchod so starožitnosťami. *Nepýtaj sa na bronzovú krysu.*

V skutočnosti sú pôvodné informácie pravdivé alebo nepravdivé rovnako ako výsledná odpoveď – odrážajú skutočnosť a možno ich posudzovať pomocou porovnania so skutočnosťou. Ak si napríklad myslíte, že 920 žien z 10 000 v danej vzorke má rakovinu prsníka, ale skutočné číslo je 100 z 10 000, potom sú vaše pôvodné údaje nesprávne. Pre náš konkrétny problém mohli byť pôvodné údaje určené troma štúdiami – štúdiou záznamov prípadov žien s rakovinou prsníka, aby sa ukázalo, koľké z nich dostalo pozitívny výsledok mamografie, štúdiou žien bez rakoviny prsníka, aby sa ukázalo, koľko z nich dostalo pozitívny výsledok mamografie, a epidemiologickou štúdiou o výskyte rakoviny prsníka v konkrétnej demografickej skupine.

* * *

Pravdepodobnosť p(A & B) je to isté ako p(B & A), ale p(A | B) nie je to isté ako p(B | A), a p(A & B) je celkom iné ako p(A | B). Je častou chybou mýliť si niektoré alebo všetky z týchto hodnôt.

Aby se sa zoznámili so všetkými týmito hodnotami a vzťahmi medzi nimi, zahráme sa „sleduj stupne voľnosti“. Napríklad dve hodnoty p(rakovina) a p(~rakovina) majú spolu iba 1 stupeň voľnosti kvôli všeobecnému zákonu p(A) + p(~A) = 1. Ak viete, že p(~rakovina) = 0,99, viete zistiť p(rakovina) = 1 – p(~rakovina) = 0,01.

Hodnoty p(pozitívny | rakovina) a p(~pozitívny | rakovina) majú spolu tiež iba jeden stupeň voľnosti; buď nejaká žena s rakovinou prsníka dostane pozitívny mamogram alebo nedostane. Na druhej strane, p(pozitívny | rakovina) a p(pozitívny | ~rakovina) majú *dva* stupne voľnosti. Môžete mať mamografický test, ktorý vráti pozitívny výsledok pre 80 % pacientok s rakovinou a 9,6 % zdravých pacientok, alebo ktorý vráti pozitívny výsledok pre 70 % pacientok s rakovinou a 2 % zdravých pacientok, alebo dokonca test zdravia, ktorý vráti „pozitívny“ výsledok pre 30 % pacientok s rakovinou a 92 % zdravých pacientok. Tieto dve hodnoty, výstupy mamografického testu pre pacientky s rakovinou a

výstupy mamografického testu pre zdravé pacientky, sú z matematického hľadiska nezávislé; jedno sa nedá nijako vypočítať z druhého, takže majú spolu dva stupne voľnosti.

Čo p(pozitívny & rakovina), p(pozitívny | rakovina) a p(rakovina)? Máme tu tri hodnoty; koľko majú stupňov voľnosti? V tomto prípade musí platiť rovnica

p(pozitívny & rakovina) = p(pozitívny | rakovina) × p(rakovina).

Táto rovnica zmenšuje počet stupňov voľnosti o jeden. Ak poznáme podiel pacientok s rakovinou, a pravdepodobnosť, že pacientka s rakoviou má pozitívny mamogram, môžeme z toho odvodiť podiel pacientok, ktoré majú rakovinu prsníka *a* pozitívny mamogram vynásobením.

Podobne, ak poznáme počet pacientok s rakovinou prsníka a pozitívnym mamogramom, aj počet pacientok s rakovinou prsníka, vieme odhadnúť pravdepodobnosť, že žena s rakovinou prsníka dostane pozitívny mamogram vydelením: p(pozitívny | rakovina) = p(pozitívny & rakovina) / p(rakovina). V skutočnosti sa presne takto kalibrujú takéto lekárske testy; urobíte štúdiu na 8520 ženách s rakovinou prsníka a vidíte, že je tam 6816 (plusmínus) žien s rakovinou prsníka a pozitívnym mamogramom, potom vydelíte 6816 / 8520 a zistíte, že 80 % žien s rakovinou prsníka malo pozitívny mamogram. (Zhodou okolností, ak omylom vydelíte 8520 / 6816 namiesto naopak, vaše výpočty začnú robiť zvláštne veci, napríklad trvať na tom, že 125 % žien s rakovinou prsníka a pozitívnym mamogramom má rakovinu prsníka. Toto je podľa mojej skúsenosti častá chyba pri robení bayesovskej aritmetiky.) A nakoniec, ak viete p(pozitívny & rakovina) a p(pozitívny | rakovina), viete si odvodiť, koľko pacientok s rakovinou muselo byť pôvodne. Tieto tri údaje majú teda spolu dva stupne voľnosti; ak poznáme dva z nich, vieme si odvodiť tretí.

A čo p(pozitívny), p(pozitívny & rakovina) a p(pozitívny & ~rakovina)? Opäť, tieto tri premenné majú iba dva stupne voľnosti. Rovnica, ktorá nám berie jeden stupeň voľnosti je

p(pozitívny) = p(pozitívny & rakovina) + p(pozitívny & ~rakovina).

Takto sa vlastne počíta p(pozitívny); zistíme počet žien s rakovinou prsníka, ktoré majú pozitívny mamogram, a počet žien bez rakoviny prsníka, ktoré majú pozitívy mamogram, potom ich sčítame a dostaneme celkový počet žien s pozitívnym mamogramom. Bolo by čudné ísť robiť štúdiu na zistenie počtu žien s pozitívnym mamogramom – iba toto jediné číslo a nič viac – ale teoreticky by ste to mohli urobiť. A keby ste potom urobili ďalšiu štúdiu a zistili počet žien, ktoré mali pozitívny mamogram *a* rakovinu prsníka, poznali by ste zároveň počet žien s pozitívnym mamogramom a *bez* rakoviny prsníka – žena s pozitívnym mamogramom buď má rakovinu prsníka alebo nie. Vo všeobecnosti p(A & B) + p(A & ~B) = p(A). Symetricky, p(A & B) + p(~A & B) = p(B).

Čo p(pozitívny & rakovina), p(pozitívny & ~rakovina), p(~pozitívny & rakovina) a p(~pozitívny & ~rakovina)? Mohli by ste byť zo začiatku v pokušení myslieť si, že tieto štyri hodnoty majú iba dva stupne voľnosti – že môžete napríklad získať p(pozitívny & ~rakovina) vynásobením p(pozitívny) × p(~rakovina), a že takto všetky štyri údaje získate iba pomocou dvoch údajov p(pozitívny) a p(rakovina). Ale nie je to tak! p(pozitívny & ~rakovina) = p(pozitívny) × p(~rakovina) iba ak sú obe pravdepodobnosti *štatisticky nezávislé* – ak šanca, že žena má rakovinu nejako nesúvisí s tým, či má pozitívny mamogram. Ako si spomínate, toto vyžaduje, aby sa dve podmienené pravdepodobnosti rovnali – požiadavka, ktorá by odstránila jeden stupeň voľnosti. Ak si pamätáte, že tieto štyri čísla sú skupiny A, B, C a D, môžete sa pozrieť na tieto skupiny a uvedomiť si, že teoreticky môžete dať do ľubovoľnej z týchto štyroch skupín ľubovoľný počet ľudí. Ak začnete si skupinou 80 žien s rakovinou prsníka a pozitívnym mamogramom, nie je dôvod, prečo by ste nemohli pridať ďalšiu skupinu 500 žien s rakovinou prsníka a negatívnym mamogramom, potom skupinu 3 žien bez rakoviny prsníka a s negatívnym mamogramom, a tak ďalej. Takže teraz sa zdá, že tieto štyri údaje majú štyri stupne voľnosti. Aj by mali, okrem toho, že ak ich vyjadrujeme ako *pravdepodobnosti*, potrebujeme ich normalizovať na *zlomky* z celkovej skupiny, čo pridáva podmienku, že p(pozitívny & rakovina) + p(pozitívny & ~rakovina) + p(~pozitívny & rakovina) + p(~pozitívny & ~rakovina) = 1. Táto rovnica uberá jeden stupeň voľnosti, čím nám necháva tri stupne voľnosti pre štyri hodnoty. Ak máte dané *percentá* žien v skupinách A, B a D, môžete z toho odvodiť percento žien v skupine C.

Ak sú dané štyri skupiny A, B, C a D, je veľmi priamočiare spočítať všetko ostatné:

p(rakovina) = A + B / (A + B + C + D)

p(~pozitívny | rakovina) = B / (A + B)

a tak ďalej. Keďže { A, B, C, D } obsahuje tri stupne voľnosti, vyplýva z toho, že celá množina pravdepodobností spájajúca mieru rakoviny a výsledky testu obsahuje iba tri stupne voľnosti. Pamätajte, že v našich problémoch vždy potrebujeme *tri* informácie – pôvodnú pravdepodobnosť a dve podmienené pravdepodobnosti – ktoré majú dokopy tri stupne voľnosti. V skutočnosti by pri bayesovských problémoch *ľubovoľné* tri hodnoty s troma stupňami voľnosti mali logicky špecifikovať celý problém.

* * *

Pravdepodobnosť, že nejaký test dáva správny pozitívny výsledok, delená *pravdepodobnosťou, že tento test dáva falošný pozitívny výsledok*, sa označuje ako *pomer pravdepodobností* daného testu. Vyjadruje pomer pravdepodobností lekárskeho testu všetko, čo možno vedieť o užitočnosti tohto testu?

Nie, nevyjadruje! Pomer pravdepodobností vyjadruej všetko, čo možno vedieť o *význame pozitívneho* výsledku lekárskeho testu, ale nie je známy význam *negatívneho* výsledku, ani frekvencia s akou je tento test užitočný. Napríklad mamografia s mierou úspešnosti 80 % pre pacientky s rakovinou prsníka a mierou falošných pozitívnych výsledkov 9,6 % pre zdravé pacientky má rovnaký pomer pravdepodobností ako test s mierou úspešnosti 8 % a mierou falošných negatívnych výsledkov 0,96 %. Hoci oba tieto testy majú rovnaký pomer pravdepodobností, prvý test je v každom ohľade užitočnejší – častejšie nájde chorobu, a negatívny výsledok je silnejšou indíciou zdravia.

* * *

Predstavte si, že použijete *dva* testy na rakovinu prsníka po sebe – povedzme štandardnú mamografiu a ešte nejaký iný test, ktorý je *nezávislý* od mamografie. Keďže neviem o žiadnom takom teste, ktorý by bol závislý od mamografie, vymyslím si jeden za účelom tohto problému, a nazvem ho Tamsov-Braylorov Test Delenia, ktorý kontroluje, či sa nejaké bunky delia rýchlejšie než iné bunky. Budeme predpokladať, že Tams-Braylor dáva pravý pozitívny výsledok pre 90 % pacientok s rakovinou, a dáva falošný pozitívny výsledok pre 5 % pacientok bez rakoviny. Predpokladajme, že výskyt rakoviny prsníka je 1 %. Ak pacientka dostane pozitívny výsledok aj na mamograme aj na Tams-Braylorovi, aká je výsledná pravdepodobnosť, že má rakovinu prsníka?

Jeden spôsob riešenia tohto problému by bol vziať výslednú pravdepodobnosť po pozitívnej mamografii, ktorú sme už spočítali ako 7,8 %, a zadať toto do Tamsovho-Braylorovho testu ako novú pôvodnú pravdepodobnosť. Ak to urobíme, zistíme, že vychádza výsledok 60 %.

Predpokladajme, že pôvodný výskyt rakoviny prsníka v populácii je 1 %. Predpokladajme, že ako lekári máme repertoár troch nezávislých testov na rakovinu prsníka. Náš prvý test A je mamografia, ktorá má pomer podmienených pravdepodobností 80 % / 9,6 % = 8,33. Druhý test B má pomer podmienených pravdepodobností 18,0 (napríklad 90 % verzus 5 %); a tretí test C má pomer pravdepodobností 3,5 (čo by mohlo byť 70 % verzus 20 %, alebo 35 % verzus 10 %; to je jedno). Predpokladajme, že pacientka dostala pozitívny výsledok na všetkých troch testoch. Aká je pravdepodobnosť, že pacientka má rakovinu prsníka?

Tu je trik na zjednodušenie účtovníctva. Ak je pôvodný výskyt rakoviny prsníka v demografickej skupine 1 %, potom 1 zo 100 žien má rakovinu prsníka, a 99 zo 100 žien nemá rakovinu prsníka. Ak teda prepíšeme *pravdepodobnosť* 1 % ako *pomer šancí*, šance sú 1 : 99.

A pomery podmienených pravdepodobností troch testov A, B a C sú:

```
  8,33 : 1 = 25 : 3
  18,0 : 1 = 18 : 1
   3,5 : 1 = 7 : 2
```

Šanca, že žena s rakovinou prsníka dostane pozitívny výsledok vo všetkých troch testoch, verzus že žena bez rakoviny prsníka dostane pozitívny výsledok vo všetkých troch testoch, sa rovná:

1 × 25 × 18 × 7 : 99 × 3 × 1 × 2 = 3 150 : 592

Aby sme zo šancí opäť dostali pravdepodobnosti, napíšeme iba:

3 150 / (3 150 + 594) = 84 %

Toto funguje vždy bez ohľadu na to, ako sú zapísané pomery šancí, napríklad 8,33 : 1 je to isté ako 25 : 3 alebo 75 : 9. Nezáleží na tom, v akom poradí sa testy aplikujú, alebo v akom poradí počítame výsledky. Dôkaz ponechávam ako cvičenie pre čitateľa.

* * *

E.T. Jaynes v *Pravdepodobnostná teória s aplikáciami vo vede a inžinierstve* odporúča, aby sa dôveryhodnosť a dôkazy merali v decibeloch.[177]

V decibeloch?

Decibely sa používajú na meranie exponenciálnych rozdielov v intenzite. Napríklad ak má zvuk automobilového klaksóna 10 000-krát viac energie (na štvorcový meter za sekundu) než zvuk budíka, klaksón automobilu je o 40 decibelov hlasnejší. Zvuk vtáčieho spevu môže mať 1 000-krát menej energie než budík, a teda je o 30 decibelov tichší. Aby ste dostali počet decibelov, vezmete logaritmus pri základe 10 a vynásobíte ho 10.

decibely = 10 × log 10 (intenzita)

alebo

intenzita = $10^{\text{decibely / 10}}$

Predpokladajme, že začneme s pôvodnou pravdepodobnosťou 1 %, že žena má rakovinu prsníka, čo zodpovedá pomeru šancí 1 : 99. A potom aplikujeme tri testy s pomermi podmienených pravdepodobností 25 : 3, 18 : 1 a 7 : 2. *Mohli* by ste tieto čísla vynásobiť... alebo by ste mohli skrátka sčítať ich logaritmy:

```
10 × log 10 (1 / 99) = -20
10 × log 10 (25 / 3) = 9
10 × log 10 (18 / 1) = 13
 10 × log 10 (7 / 2) = 5
```

Začne to ako značne nepravdepodobné, že by nejaká žena mala rakovinu prsníka – naša úroveň dôveryhodnosti je -20 decibelov. Potom prídu tri výsledky testov, ktoré zodpovedajú 9, 13 a 5 decibelom indície. To zdvihne celkovú úroveň dôveryhodnosti spolu o 27 decibelov, čo znamená, že pôvodná dôveryhodnosť -20 decibelov ide na výslednú dôveryhodnosť 7 decibelov. Šance teda idú z 1 : 99 na 5 : 1, a pravdepodobnosť ide z 1 % na 83 %.

* * *

> Ste mechanik zariadení. Keď zariadenie prestane fungovať, že to v 30 % prípadov kvôli upchanej hadici. Ak je hadica na zariadení upchatá, je pravdepodobnosť 45 %, že po štuchnutí zo zariadenia vyletia iskry. Ak hadica na zariadení nie je upchatá, je pravdepodobnosť iba 5 %, že po štuchnutí zo zariadenia vyletia iskry. Zákazník vám priniesol nefungujúce zariadenie. Štuchnete do zariadenia a zistíte, že z neho lietajú iskry. Aká je pravdepodobnosť, že toto iskriace zariadenie má upchatú hadicu?

Aká je postupnosť aritmetických operácií, ktoré ste robili pri riešení tohto problému?

(45 % × 30 %) / (45 % × 30 % + 5 % × 70 %)

Podobne, aby sme našli šancu, že žena s pozitívnym mamogramom ná rakovinu prsníka, počítali sme:

```
    p(pozitívny | rakovina) × p(rakovina)
-----------------------------------------------
(    p(pozitívny | rakovina) × p(rakovina)    )
(  + p(pozitívny | ~rakovina) × p(~rakovina)  )
```

čo je

```
p(pozitívny & rakovina)
-------------------------------------------------
p(pozitívny & rakovina) + p(pozitívny & ~rakovina)
```

čo je

```
p(pozitívny & rakovina)
-----------------------
p(pozitívny)
```

čo je

177 Edwin T. Jaynes, „Probability Theory, with Applications in Science and Engineering,“ [Pravdepodobnostná teória s aplikáciami vo vede a inžinierstve] Unpublished manuscript (1974).

p(rakovina | pozitívny)

Celkom všeobecný tvar tohto výpočtu sa nazýva *Bayesova veta* alebo *Bayesovo pravidlo*:

Bayesova veta:

$$p(A \mid X) = \frac{p(X \mid A) \times p(A)}{p(X \mid A) \times p(A) + p(X \mid \sim A) \times p(\sim A)}$$

Ak je daný jav A, ktorý chceme preskúmať, a pozorovanie X, ktoré je indíciou ohľadom A – napríklad v predchádzajúcom príklade A je rakovina prsníka a X je pozitívny mamogram – Bayesova veta nám hovorí, ako by sme mali *aktualizovať* našu pravdepodobnosť A pri *novej indícii* X.

V tomto bode môže Bayesova veta vyzerať očividne samozrejmá alebo priam tautologická, a nie vzrušujúca a nová. Ak je to tak, tento úvod *celkom splnil* svoj cieľ.

* * *

Bayesova veta opisuje, čo tvorí „indíciu“ a koľko indície to je. Štatistické modely posudzujeme porovnaním s *bayesovskou metódou*, pretože v štatistike je bayesovská metóda to najlepšie možné – bayesovská metóda definuje maximálne množstvo výsledku, ktoré možno vyťažiť z daného kusu indície, rovnako ako termodynamika definuje maximálne množstvo práce, ktorú možno vyžažiť z teplotného rozdielu. To je dôvod, prečo počujete kognitívnych vedcov hovoriť o *bayesovskom uvažovaní*. V kognitívnej vede je *bayesovské uvažovanie* technicky presný pojem, ktorým označujeme *rozumnú myseľ*.

Existuje aj veľa všeobecných heuristík o ľudskom rozmýšľaní, ktoré sa môžete naučiť z pohľadu na Bayesovu vetu.

Napríklad, v mnohých diskusiách o Bayesovej vete môžete počuť, ako kognitívni psychológovia hovoria, že ľudia *dostatočne nezohľadňujú pôvodné frekvencie*, čím myslia, že keď sa ľudia stretnú v problémom, kde nejaká indícia X naznačuje, že by mohla platiť podmienka A, majú sklon posudzovať pravdepodobnosť A iba podľa toho, ako veľmi sa zdá, že indícia X zodpovedá A, bez zváženia pôvodnej frekvencie A. Ak si napríklad myslíte v príklade o mamografii, že pravdepodobnosť, že daná žena má rakovinu prsníka je v rozsahu 70 % - 80 %, tento druh uvažovania je necitlivý na pôvodnú frekvenciu zadanú v probléme; nevšíma si, či 1 % žien alebo 10 % žien má na začiatku rakovinu prsníka. „Viac si všímajte pôvodnú frekvenciu!“ je jedna z mnohých vecí, na ktoré ľudia potrebujú stále pamätať, aby tým čiastočne kompenzovali naše zabudované nedostatky.

Súvisiaca chyba je venovať priveľa pozornosti p(X | A) a nie dosť p(X | ~A) pri určovaní, nakoľko je X indíciou pre A. Stupeň, nakoľko je výsledok X *indíciou pre* A závisí nielen od sily tvrdenia, *že by sme sme očakávali výsledok X, keby A bola pravda*, ale aj od sily tvrdenia, *že by sme* ***neočakávali*** *výsledok X keby A nebola pravda*. Napríklad, keď prší, veľmi silno to naznačuje, že tráva je mokrá – p(mokrá tráva | prší) ~ 1 – ale videnie mokrej trávy nemusí nutne znamenať, že práve pršalo; možno bol zapnutý zavlažovač alebo sa pozeráta na rannú rosu. Keďže p(mokrá tráva | ~prší) je podstatne vyššia než nula, p(prší | mokrá tráva) je podstatne nižšia než jedna. Na druhej strane, keby tráva *nikdy* nebola mokrá keď neprší, potom poznanie, že tráva je mokrá, by *vždy* ukazovalo, že pršalo, p(prší | mokrá tráva) ~ 1, dokonca aj keby p(mokrá tráva | prší) = 50 %; čiže dokonca aj keby tráva bola mokrá iba v 50 % prípadov, keď pršalo. Indícia je vždy výsledkom *rozdielu* medzi dvoma podmienenými pravdepodobnosťami. *Silná* indícia nie je výsledok veľmi vysokej pravdepodobnosti, že A vedie k X, ale výsledkom veľmi *nízkej* pravdepodobnosti, že *nie-A* mohlo viesť k X.

Bayeovská revolúcia vo vede je poháňaná nielen tým, že si stále viac kognitívnych vedcov zrazu všíma, že mentálne javy majú v sebe bayesovskú štruktúru; nielen tým, že sa vedci v každom odvetví učia hodnotiť svoje štatistické metódy pomocou porovnania s bayesovskou metódou; ale preto, lebo myšlienka *samotnej vedy je špeciálnym prípadom Bayesovej vety; experimentálna indícia je bayesovská indícia*. Bayesovskí revolucionári tvrdia, že keď robíte experiment a dostanete indíciu, ktorá „potvrdzuje“ alebo „vyvracia“ vašu teóriu, toto potvrdenie a toto vytvrátenie sa riadi bayesovskými pravidlami. Napríklad musíte zohľadniť nielen to, či vaša teória predpovedá daný jav, ale aj či všeky zvyšné možno vysvetlenia predpovedajú tento jav.

Predtým bol najobľúbenejšou teóriu vedy asi *falzifikacionizmus* Karla Poppera – to je stará filozofia, ktorú v súčasnosti bayesovská revolúcia zvrháva z trónu. Myšlienka Karla Poppera, že teórie možno definitívne vyvrátiť, ale nikdy nie definitívne potvrdiť, je ďalším špeciálnym prípadom bayesovských pravidiel; ak p(X | A) ~ 1 – ak teória robí jasnú predpoveď – potom pozorovanie ~X silno vyvracia A. Na druhej strane, ak p(X | A) ~ 1, a my pozorujeme X, toto definitívne nepotvrdzuje danú teóriu; môže existovať nejaká iná okolnosť B taká, že p(X | B) ~ 1, a v takom prípade pozorovanie X neuprednostňuje A pred B. Aby sme pozorovaním X definitívne potvrdili A, museli by sme vedieť nielen, že p(X | A) ~ 1, ale že p(X | ~A) ~ 0, čo je niečo, čo nemôžeme vedieť, pretože nemôžeme vymenovať všetky možné alternatívne vysvetlenia. Napríklad keď Einsteinova teória všeobecnej relativity zvrhla Newtonovu neuveriteľne dobre potvrdenú teóriu gravitácie, ukázalo sa, že všetky Newtonove predpovede boli iba špeciálnym prípadom Einsteinových predpovedí.

Popperovu filozofiu môžete dokonca matematicky formalizovať. Pomer podmienených pravdepodobností pre X, p(X | A) / p(X | ~A), určuje, nakoľko pozorovanie X posúva pravdepodobnosť A; pomer podmienených pravdepodobností určuje, *aká silná* indícia je X. Takže, vo vašej teórii A môžete predpovedať X s pravdepodobnosťou 1, ak chcete; nemôžete však ovládať menovateľ pomeru pravdepodobností, p(X | ~A) – vždy budú existovať nejaké alternatívne teórie, ktoré tiež predpovedajú X, a hoci my používame tú najjednoduchšiu teóriu, ktorá je v súlade so súčasnými indíciami, jedného dňa môžeme nájsť nejakú indíciu, ktorú nejaká alternatívna teória predpovedá, ale vaša teória nie. To je skrytý háčik, ktorý zvrhol Newtonovu teóriu gravitácie. Existuje teda hranica, koľko výsledku môžete dostať z úspešných predpovedí; existuje hranica, ako vysoko ide pomer podmienených pravdepodobností pre *potvrdzujúce* indície.

Na druhej strane, ak natrafíte na nejaký kúsok indície Y, ktorú vaša teória jednoznačne *nepredpovedá*, toto je *ohromne* silná indícia proti vašej teórii. Ak je p(Y | A) mikroskopické, potom aj pomer podmienených pravdepodobností bude mikroskopický. Napríklad ak p(Y | A) je 0,0001 % a p(Y | ~A) je 1 %, potom pomer podmienených pravdepodobností p(Y | A) / p(Y | ~A) bude 1 : 10000. -40 decibelov indície! Alebo, ak otočíme pomer podmienených pravdepodobností, ak je p(Y | A) *veľmi malé*, potom p(Y | ~A) / p(Y | A) bude veľmi veľké, čo znamená, že pozorovanie Y výrazne uprednostní ~A pred A. Falzifikácia je omnoho silnejšia než potvrdenie. To je dôsledkom predchádzajúcej pointy, že *veľmi silná* indícia nie je výsledkom veľmi vysokej pravdepodobnosti, že A vedie k X, ale výsledkom veľmi *malej* pravdepodobnosti, že *nie-A* by mohlo viesť k X. Toto je presné bayesovské pravidlo, na ktorom sa zakladá heuristická hodnota Popperovho falzifikacionizmu.

Podobne, Popperov výrok, že myšlienka musí byť falzifikovateľná, možno vysvetliť ako prejav bayesovského zákona zachovania pravdepodobnosti; ak je výsledok X pozitívnou indíciou pre danú teóriu, potom by výsledok ~X mal túto teóriu do istej miery vyvracať. Ak sa pokúšate interpretovať aj X aj ~X ako „potvrdenie“ teórie, bayesovské pravidlá hovoria, že toto sa nedá! Aby ste zvýšili pravdepodobnosť teórie, *musíte* ju vystaviť testom, ktoré môžu potenciálne znížiť jej pravdepodobnosť; toto nie je iba pravidlo na odhaľovanie rádoby podvodníkov v spoločenskom procese vedy, ale dôsledkom bayesovskej teórie pravdepodobnosti. Na jednej strane, Popperova myšlienka, že existuje *iba* falzifikácia a neexistuje *nič také* ako potvrdenie, sa ukazuje nesprávna. Bayesova veta ukazuje, že falzifikácia je *veľmi silná* indícia v porovnaní s potvrdením, ale falzifikácia je stále vo svojej podstate pravdepodobnostná; neriadi sa principiálne inými pravidlami než potvrdenie, ako tvrdil Popper.

Zisťujeme teda, že mnohé javy v kognitívnych vedách, plus štatistické metódy používané vedcami, plus samotná vedecká metóda, sa všetky ukazujú ako špeciálne prípady Bayesovej vety. Preto bayesovská revolúcia.

* * *

Keď už sme si Bayesovu vetu uviedli výslovne, môžeme výslovne diskutovať o jej častiach.

$$p(A \mid X) = \frac{p(X \mid A) \times p(A)}{p(X \mid A) \times p(A) + p(X \mid \sim A) \times p(\sim A)}$$

Začneme s p(A | X). Ak niekedy zistíte, že sa pletiete ohľadom toho, čo je A a čo je X v Bayesovej vete, začnite p(A | X) na ľavej strane rovnice; to je najjednoduchšia časť na vysvetlenie. A je to, čo chceme vedieť. X je ako ju pozorujeme; X je indícia, ktorú používame na robenie úsudkov o A. Pamätajte, že pri každom výraze p(Q | P) chceme vedieť pravdepodobnosť Q pri danom P, stupeň nakoľko P implikuje Q – zmysluplnejší zápis, na ktorého zavedenie je už prilíš neskoro, by bol p(Q <- P).

p(Q | P) úzko súvisí s p(Q & P), ale nie je to to isté. Vyjadrené ako pravdepodobnosť alebo zlomok, p(Q & P) je podiel vecí, ktoré majú vlastnosť Q a vlastnosť P v rámci všetkých vecí; čiže podiel „žien s rakovinou prsníka a pozitívnym mamogramom“ v rámci skupiny všetkých žien. Ak je celkový počet žien 10 000 a 80 žien má rakovinu prsníka a pozitívny mamogram, potom p(Q & P) je 80 / 10 000 = 0,8 %. Môžete povedať, že absolútna hodnota 80 sa normalizuje na pravdepodobnosť relatívne voči skupine všetkých žien. Alebo aby to bolo jasnejšie, predpokladajme, že máme skupinu 641 žien s rakovinou prsníka a pozitívnym mamogramom v celkovej vzorke 89 031 žien. 641 je absolútna hodnota. Ak si vyberiete z celej vzorky náhodnú ženu, potom pravdepodobnosť, že ste vybrali ženu s rakovinou prsníka a pozitívnym mamogramom je p(Q & P), čiže 0,72 % (v tomto prípade).

Na druhej strane, p(Q | P) je podiel vecí, ktoré majú vlastnosť Q a vlastnosť P v rámci *všetkých vecí, ktoré majú* P; čiže podiel žien s rakovinou prsníka a pozitívnym mamogramom v rámci skupiny *všetkých žien s pozitívnym mamogramom*. Ak je 641 žien s rakovinou prsníka a pozitívnym mamogramom, 7 915 žien s pozitívnym mamogramom, a 89 031 žien, potom p(Q & P) je pravdepodobnosť vybrania jednej z týchto 641 žien, ak vyberáte náhodnú ženu z celej skupiny 89 031, zatiaľ čo p(Q | P) je pravdepodobnosť vybrania jednej z týchto 641 žien, ak vyberáte náhodnú ženu z menšej skupiny 7 915.

V istom zmysle p(Q | P) v skutočnosti znamená p(Q & P | P), ale písať zakaždým to P navyše by bolo zbytočné opakovanie. Už *viete*, že to má vlastnosť P, takže *skúmate* vlastnosť Q – aj keď sa pozeráte na veľkosť skupiny Q & P v rámci skupiny P, nie na veľkosť skupiny Q v rámci skupiny P (čo by bol nezmysel). Toto znamená povedať, že vlastnosť na pravej strane berieme ako danú; znamená to, že viete, že pracujete iba v rámci skupiny vecí, ktoré majú vlastnosť P. Keď obmedzíte svoju pozornosť tak, že vidíte iba túto menšiu skupinu, mnoho iných pravdepodobností sa zmení. Ak beriete P ako *dané*, potom p(Q & P) sa rovná skrátka p(Q) – prinajmenšom *vzhľadom na skupinu* P. *Stará* frekvencia p(Q) „veci, ktoré majú vlastnosť Q v rámci celej vzorky“ sa upravila na novú frekvenciu „veci, ktoré majú vlastnosť Q v rámci podskupiny vecí, ktoré majú vlastnosť P“. Ak je P *dané*, ak je P celý náš svet, potom pozerať na Q & P je to isté ako pozerať iba na Q.

Ak obmedzíte svoju pozornosť iba na populáciu vajec, ktoré sú namaľované na modro, potom sa zrazu „pravdepodobnosť, že vajce obsahuje perlu“ stane iným číslom; tento podiel je iný pre populáciu modrých vajec než pre populáciu všetkých vajec. Čo je *dané*, tá vlastnosť, ktorá obmedzuje našu pozornosť, je vždy na *pravej* strane p(Q | P); toto P sa stáva naším svetom, všetkým, čo vidíme, a naopak toto „dané“ P má vždy pravdepodobnosť 1 – to znamená povedať, že P je dané. Takže p(Q | P) znamená: „ak P má pravdepodobnosť 1, akú pravdepodobnosť má Q?“ alebo „ak obmedzíme našu pozornosť iba na tie veci alebo udalosti, kde P je pravda, aká je pravdepodobnosť Q?“ Q naopak *nie je* dané, nie je isté – jeho pravdepodobnosť môže byť 10 % alebo 90 % alebo hocijaké iné číslo. Ak teda použijete Bayesovu vetu a napíšete časť na ľavej strane ako p(A | X) – ako *aktualizovať* pravdepodobnosť A po videní X, nová pravdepodobnosť A pri *danom* X, stupeň nakoľko X *implikuje* A – môžete povedať, že X je vždy *pozorovanie* alebo *indícia*, a A je vlastnosť, ktorá sa vyšetruje, to, o čom chcete vedieť.

* * *

Pravú stranu Bayesovej vety odvodíme z ľavej strany pomocou týchto krokov:

$$p(A \mid X) = p(A \mid X)$$

$$p(A \mid X) = \frac{p(X \,\&\, A)}{p(X)}$$

$$p(A \mid X) = \frac{p(X \& A)}{p(X \& A) + p(X \& \sim A)}$$

$$p(A \mid X) = \frac{p(X \mid A) \times p(A)}{p(X \mid A) \times p(A) + p(X \mid \sim A) \times p(\sim A)}$$

Keď je odvodzovanie dokončené, všetky implikácie na pravej strane rovníce sú v tvare p(X | A) alebo p(X | ~A), zatiaľ čo implikácia na ľavej strane je p(A | X). Táto symetria vzniká preto, lebo základné *kauzálne vzťahy* sú vo všeobecnosti implikáciami od faktov k pozorovaniam, čiže od rakoviny prsníka po pozitívny mamogram. Základné *kroky v uvažovaní* sú vo všeobecnosti implikáciami od pozorovaní k faktom, čiže od pozitívneho mamogramu po rakovinu prsníka. Ľavá strana Bayesovej vety je základný krok *odvodenia* od pozorovania pozitívneho mamogramu po záver, že je zvýšená pravdepodobnosť rakovinu prsníka. Implikácia je zapísaná sprava doľava, čiže na ľavú stranu rovnice píšeme p(rakovina | pozitívny). Pravá strana Bayesovej vety opisuje základné *kauzálne* kroky – napríklad od rakoviny prsníka k pozitívnemu mamogramu – a preto majú implikácie na pravej strane Bayesovej vety tvar p(pozitívny | rakovina) alebo p(pozitívny | ~rakovina).

A to je Bayesova veta. Rozumné odvodenie na ľavej strane, fyzikálna kauzalita na pravej strane; rovnica a mysľou na jednej strane a skutočnosťou na druhej. Pamätáte sa, ako sa ukázalo, že veda je špeciálnym prípadom Bayesovej vety? Keby ste to chceli vyjadriť básnicky, mohli by ste povedať, že Bayesova veta zväzuje rozmýšľanie s fyzikálnym vesmírom.

Okej, hotovo.

* →

Reverend Bayes hovorí:

Teraz ste zasvätení do bayesovskej konšpirácie.

→ http://yudkowsky.net/rational/bayes

Kniha IV.

Obyčajná skutočnosť

Svet: Úvod

(napísal Rob Bensinger)

Predchádzajúce eseje pojednávali o ľudskom rozmýšľaní, jazyku, cieľoch a spoločenskej dynamike. Matematika, fyzika a biológia boli citované na vysvetlenie vzorcov v ľudskom správaní, ale málo sa povedalo o mieste človeka v prírode, alebo o prirodzenom svete ako takom.

Rovnako ako bolo užitočné postaviť ľudí ako *systémy zamerané na cieľ* do kontrastu s nie ľudskými procesmi v evolučnej biológii a umelej inteligencii, v nasledujúcich postupnostiach esejí bude užitočné postaviť ľudí ako *fyzikálne systémy* do kontrastu s nie ľudskými procesmi, ktoré *nie sú* mysle.

My ľudia sme napokon postavení z častí, ktoré nie sú ľuďmi. Svet atómov vôbec nevyzerá ako ten svet, o ktorom bežne rozmýšľame, a rozhodne vôbec nevyzerá ako vedomí obyvatelia sveta, o ktorých bežne rozmýšľame. Ako povedal Giulio Giorello v rozhovore s Danielom Dennettom: „Áno, máme dušu. Ale tá sa skladá z mnohých drobných robotov.“[178]

178 Daniel C. Dennett, *Freedom Evolves* (Viking Books, 2003).

Obyčajná skutočnosť obsahuje sedem postupností esejí na túto tému. Prvé tri predkladajú otázku, aký je vzťah medzi svetom ľudí a svetom, ktorý odhalila fyzika: „Zákony pravdy“ (o základných súvislostiach medzi fyzikou a ľudským myslením), „Redukcionizmus pre začiatočníkov“ (o projekte vedeckého vysvetľovania javov), a „Radosť z púhej skutočnosti“ (o emocionálnom, osobnom význame vedeckého svetonázoru). Za tým nasledujú dve postupnosti, ktoré idú do väčšej hĺbky ohľadom konkrétnych akademických debát: „Fyzikalizmus pre mierne pokročilých“ (o ťažkom probléme vedomia) a „Kvantová fyzika a mnohé svety“ (o probléme merania vo fyzike). Napokon postupnosť „Veda a rozumnosť“ a esej Technické vysvetlenie technického vysvetlenia spájajú tieto myšlienky dokopy a dávajú ich do vzťahu k vedeckej praxi.

Diskusie o vedomí a kvantovej fyzike ilustrujú relevantnosť redukcionizmu pre súčasné kontroverzie vo vede a filozofii. Pre tých, ktorý chcú vedieť trochu viac kontextu, tu poviem pár slov navyše na tieto dve témy. Pre tých, ktorí to chcú preskočiť: preskakujte!

Mysle vo svete

Môžeme vôbec niekedy vedieť, aké to je, byť netopier?

Iste si môžeme vyvinúť lepšie kognitívne modely na predpovedanie správania netopiera, alebo jemnejšie modely neurológie netopiera – ale nie je zrejmé, že by nám toto povedalo, aký je subjektívny pocit z echolokácie, alebo aké je lietanie, *z pohľadu netopiera*.

Veru, zdá sa, akoby sme si nikdy nemohli byť ani len istí, že vôbec *existuje* niečo, čo je také ako byť netopierom. Prečo by nevedomý automat nemohol replikovať všetko vonkajšie správanie vedomého činiteľa s ľubovoľnou presnosťou? (Filozofi nazývajú takéto automaty „zombie“, hoci majú málo spoločné so zombiami z ľudovej slovesnosti – ktoré sa *celkom viditeľne* odlišujú od vedomých činiteľov!)

Rasa mimozemských psychológov by natrafila na ten istý problém pri snahe modelovať *ľudské* vedomie. Mohli by dospieť k dokonalému modelu predpovedajúcemu, čo povieme a urobíme, keď uvidíme červenú ružu, ale to neznamená, že by títo mimozemšťania plne pochopili, ako nám červenosť pripadá „vo vnútri“.

Používajúc príklady ako sú tieto, filozofi ako Thomas Nagel a David Chalmers argumentovali, že kognitívne a neurónové modely v tretej osobe nikdy nedokážu plne zachytiť vedomie v prvej osobe.[179,180] Bez ohľadu na to, ako veľa vieme o nejakom fyzikálnom systéme, je vždy logicky možné, podľa tohto názoru, že daný systém nemá žiadne zážitky v prvej osobe. Tradičný dualizmus, v ktorom sa nehmotné duše voľne vznášali okolo porušujúc fyzikálne zákony, môže byť nepravdivý; ale Chalmers trvá na slabšej téze, že vedomie je „dodatočný fakt“, ktorý nemožno plne vysvetliť fyzikálnymi faktmi.

Mnoho filozofov a vedcov považuje túto myšlienkovú líniu za presvedčivú.[181] Ak cítime intuitívnu silu tohto argumentu, mali by sme prijať jeho záver a zavrhnúť fyzikalizmus?

Určite by sme tento argument nemali odmietnuť iba preto, že *znie čudne*, alebo nám pripadá nejasne nevedecký. Ale ako dopadne tento argument po *technickom* pochopení toho, ako fungujú vysvetlenia a názory? Sú nejaké ponaučenia, ktoré si môžeme odniesť z histórie vedy, alebo z nášho chápania fyzikálnych mechanizmov, na ktorých sú založené indície. „Fyzikalizmus pre mierne pokročilých“ sa vráti k tejto otázke.

Svety vo svete

179 David J. Chalmers, *The Conscious Mind: In Search of a Fundamental Theory* (New York: Oxford University Press, 1996).

180 Thomas Nagel, „What Is It Like to Be a Bat?,“ *Philosophical Review* 83, no. 4 (1974): 435–450, http://www.jstor.org/stable/2183914.

181 V prieskume anglofónnych profesionálnych filozofov 56,5 % súhlasilo s fyzikalizmom, 21,1 % súhlasilo s anti-fyzikalizmom, a 16,4 % súhlasilo s inými názormi (napríklad „neviem“). Väčšina filozofov odmieta metafyzickú možnosť Chalmersových „zombií“, ale nezhodnú sa na tom, prečo presne Chalmersov argument so zombiami zlyháva. Kirk zhŕňa súčasné postoje voči fenomenálnemu vedomiu a dáva argumenty, ktoré sa podobajú na Yudkowskeho argumenty proti možnosti poznania neredukovateľných kvalií alebo odkazovania na ne.
David Bourget and David J. Chalmers, „What Do Philosophers Believe?,“ *Philosophical Studies* (2013): 1–36.
Robert Kirk, *Mind and Body* (McGill-Queen’s University Press, 2003).

Kvantová mechanika je v súčasnosti náš najlepší matematický model vesmíru, silne potvrdený storočím testovania. Táto teória predpokladá komplexné čísla nazývané „amplitúdy pravdepodobnosti“, pretože istá konkrétna operácia (umocnenie absolútnej hodnoty čísla na druhú – Bornovo pravidlo) nám umožňuje pravdepodobnostne predpovedať javy pri malých rozmeroch a extrémnych úrovniach energie. Táto amplitúda sa mení deterministicky podľa Schrödingerovej rovnice. V tomto procese často nadobúda čudné stavy zvané „superpozície“.

Napriek tomu, keď robíme experimenty, zdá sa, že superpozície miznú bez stopy. Keď sa nepozeráme, zdá sa, že Schrödingerova rovnica zachytáva všetko, čo sa dá vedieť o dynamike fyzikálnych systémov. Keď sa však *pozeráme*, tento čistý determinizmus sa nahrádza Bornovým pravdepodobnostným pravidlom. Je to akoby sa bežné zákony fyziky náhle zastavili vždy keď urobíme „pozorovanie“. Ako to vyjadril John Stewart Bell:

> Zdá sa, že táto teória sa výhradne stará o „výsledky meraní“ a nemá čo povedať o ničom inom. Čo presne kvalifikuje nejaké fyzikálne systémy hrať rolu „merajúceho“? Vari vlnová funkcia celého sveta čakala na skok stovky miliónov rokov, dokiaľ sa neobjavil živý jednobunkovec? Alebo musela čakať o čosi dlhšie, na nejaký lepšie kvalifikovaný systém... s PhD?

Každý sa zhodne, že táto čudná zmes Schrödingerových a Bornových pravidiel sa v praxi ukázala ako postačujúca. Avšak otázka, *kedy* presne Bornovo pravidlo vstupuje do hry, a čo to celé *znamená*, vytvorila zmätok rôznych názorov na podstatu kvantovej mechaniky.

Copenhagenská škola – Niels Bohr a ďalší objavitelia kvantovej teórie – sa veľmi rýchlo rozdelila na niekoľko štandardných spôsobov hovorenia o experimentálnych výsledkoch a o čudnom formalizme používanom na ich predpovedanie. Niektorí brali zameranie teórie na „merania“ a „pozorovania“ celkom doslovne, a predpokladali, že vedomie hrá základnú rolu vo fyzikálnom zákone, zasahuje, aby spôsobilo „kolaps“ komplexných amplitúd na pozorovania. Iní na čele s Wernerom Heisenbergom presadzovali nerealistický pohľad, podľa ktorého fyzika hovorí o našich stavoch poznania, a nie o nejakej objektívnej skutočnosti. Ešte ďalšia copenhagenská tradícia, zhrnutá sloganom „drž hubu a počítaj“, varovala pred metafyzickými špekuláciami každého druhu.

Yudkowsky používa túto vedeckú kontroverziu ako testovacie pole na niektoré ústredné myšlienky z predchádzajúcich postupností: rozlišovanie medzi mapou a územím, tajomné odpovede, Bayesiánstvo, a Occamova britva. Keďže nie je fyzik – a ani ja nie som – poskytnem tu niektoré cudzie zdroje čitateľom, ktorí si chcú overiť jeho argumenty alebo sa naučiť viac o jeho príkladoch z fyziky.

Náš matematický vesmír od Tegmarka pojednáva o množstve súvisiacich myšlienok vo filozofii a fyzike.[182] Medzi Tegmarkovými novšími myšlienkami je jeho argument, že všetky konzistentné matematické štruktúry existujú, vrátane svetov, ktorých fyzikálne zákony a hraničné podmienky sa vôbec nepodobajú na tie naše. Rozlišuje medzi týmito Tegmarkovými svetmi a multiverzami vo vedecky mainstreamovejších hypotézach -– napríklad svetmi v stochastických modeloch večnej inflácie Veľkého Tresku, a Everettovou interpretáciou mnohých svetov kvantovej fyziky.

Yudkowsky podrobne pojednáva o interpretácii mnohých svetov v odpoveď na copenhagenské interpretácie kvantovej mechaniky. Mnohé svety sa v posledných desaťročiach stali veľmi obľúbené medzi fyzikmi, najmä u kozmológov. Mnoho fyzikov ich však naďalej odmieta alebo si zachováva agnosticizmus. Filozofický (najmä) mainstreamový úvod do tejto debaty dáva Albertova *Kvantová mechanika a zážitok*.[183] Pozrite si aj v *Stanfordskej encyklopédii filozofie* heslo „Meranie v kvantovej

182 Max Tegmark, *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality* (Random House LLC, 2014).

183 David Z. Albert, *Quantum Mechanics and Experience* (Harvard University Press, 1994).

teórii“[184] a ich úvod k niekoľkým z pohľadov spájaných s „mnohými svetmi“ v heslách „Everettova formulácia relatívneho stavu“[185] a „Interpretácia mnohých svetov“.[186]

Z menej teoretickej stránky, Epsteinove *Thinking Physics* je dobrý text na tréning fyzikálnych intuícií.[187] Oplatí sa pamätať na to, že rovnako ako človek môže pochopiť väčšinu kognitívnej vedy bez poznania podstaty subjektívneho vedomia, človek môže aj pochopiť väčšinu fyziky bez ustáleného názoru na definitívnu podstatu (a veľkosť!) fyzikálneho sveta.

184 Henry Krips, „Measurement in Quantum Theory,“ [Meranie v kvantovej teórii] in *The Stanford Encyclopedia of Philosophy*, Fall 2013, ed. Edward N. Zalta.

185 Jeffrey Barrett, *Everett's Relative-State Formulation of Quantum Mechanics*, ed. Edward N. Zalta, http://plato.stanford.edu/archives/fall2008/entries/qm-everett/.

186 Lev Vaidman, „Many-Worlds Interpretation of Quantum Mechanics,“ in *The Stanford Encyclopedia of Philosophy*, Fall 2008, ed. Edward N. Zalta.

187 Lewis Carroll Epstein, *Thinking Physics: Understandable Practical Reality*, 3rd Edition (Insight Press, 2009).

O: Zákony pravdy

181. Všeobecný oheň

Vo fantasy príbehu L. Spraguea de Campa: *Neúplný čarodej* (ktorý vytýčil trasu mnohým imitáciám) je hrdina, Harold Shea, prenesený z nášho vesmíru do vesmíru severskej mytológie.[188] Tento svet je založený na mágii a nie na technológii, takže prirodzene, keď sa náš hrdina pokúsi založiť oheň pomocou zápalky, ktorú si priniesol zo Zeme, zápalka sa nezapáli.

Ja viem, že je to len fantasy príbeh, ale... ako by som to povedal...

Nie.

Koncom osemnásteho storočia Antoine-Laurent de Lavoisier objavil oheň. „Čože?“ opýtate sa. „Vari sa oheň nepoužíval už státisíce rokov?“ No áno, ľudia *používali* oheň; bol horúci, jasný, oranžovo-nejaký, a dalo sa na ňom variť. Ale nikto nevedel, ako funguje. Grécki a stredovekí alchymisti si mysleli, že oheň je základná častica, jeden zo štyroch prvkov. V Lavoisierovej dobe sa táto alchymistická paradigma postupne zmenila a skomplikovala, ale oheň bol stále považovaný za základnú časticu – vo forme „flogistonu“, čo bola veľmi tajomná látka, o ktorej sa hovorilo, že vysvetľuje oheň, a tiež mnohé ďalšie alchymistické javy.

Lavoisierov veľký prínos bol v odvážení *všetkých* kúskov chemickej skladačky, pred chemickou reakciou aj po nej. Predtým sa myslelo, že nejaká chemická premena zmenila celkovú hmotnosť materiálu: Ak jemne rozdrvený antimón vystavíte slnečným lúčom sústredeným pomocou lupy, za hodinu z antimónu zostane popol, ktorý váži o jednu desatinu viac ako pôvodný antimón – napriek tomu, že horenie sprevádza únik hustého bieleho dymu. Lavoisier odvážil *všetky* zložky takýchto reakcií, vrátane vzduchu, v ktorom sa reakcia odohrávala, a zistil, že hmota ani nevznikala ani nezanikala. Ak pri spaľovaní popol nadobudol hmotnosť, zodpovedal tomu úbytok hmotnosti vzduchu.

Lavoisiser ďalej vedel, ako oddeľovať plyny, a objavil, že horiaca sviečka zmenšuje množstvo jedného dôležitého plynu, *živého vzduchu*, a vytvára iný plyn, *stály vzduch*. Dnes ich voláme *kyslík* a *kysličník uhličitý*. Keď sa *živý vzduch* vyčerpal, oheň zhasol. Potom bolo možné uhádnuť, že horenie premieňa *živý vzduch* na *stály vzduch* a palivo na popol, a že schopnosť pokračovať v tejto premene je ohraničená dostupnosťou *živého vzduchu*.

Lavoisierov návrh priamo odporoval vtedy modernej teórii flogistonu. To samotné by bolo dosť šokujúce, ale potom sa ešte ukázalo...

Aby ste vedeli oceniť, čo príde ďalej, musíte sa vžiť do mysle osemnásteho storočia. Zabudnite na DNA, ktorá bola objavená v roku 1953. Zabudnite celú bunkovú teóriu z biológie, ktorá vznikla v roku 1839. Predstavte si, že sa pozeráte na svoju ruku, ohýbate prsty... a nemáte najmenšie tušenie, ako to funguje. Poznáte anatómiu svalov a kostí, ale nikto ani netuší „čo tým hýbe“ – prečo sa sval stiahne a uvoľní, zatiaľ čo hlina stlačená do rovnakého tvaru tak proste zostane. Predstavte si, že *vaše vlastné telo* sa skladá z tajomnej, nepochopiteľnej kaše. A potom si predstavte objav...

...že ľudia pri dýchaní spotrebovávajú *živý vzduch* a vydychujú *stály vzduch*. Ľudia tiež fungujú na spaľovaní! Lavoisier odmeral, koľko tepla zvieratá (a jeho asistent Seguin) produkujú, keď cvičia, koľko *živého vzduchu* spotrebujú a koľko *stáleho vzduchu* vydýchnu. Keď zvieratá produkovali viac tepla, spotrebovávali viac *živého vzduchu* a vydychovali viac *stáleho vzduchu*. Ľudia, tak ako oheň, spotrebúvajú palivo a kyslík; ľudia, tak ako oheň, produkujú teplo a kysličník uhličitý. Zrušte ľuďom kyslík alebo palivo, a plameň zhasne.

Zápalky sa zapaľujú škrtaním vďaka fosforu - „bezpečnostné zápalky“ majú fosfor na škrtacom prúžku; dvojzložkové zápalky majú fosfor v hlavičkách. Fosfor je vysoko reaktívny; čistý fosfor v tme žiari a môže samovoľne vzbĺknuť. (Henning Brand, ktorý vyrobil čistý fosfor v roku 1669, oznámil, že objavil elementárny oheň.) Fosfor sa teda výborne hodí aj na svoju rolu v *adenozíntrifosfáte*, ATP, hlavnej metóde vášho tela na ukladanie chemickej energie. ATP je občas označovaný ako „molekulárne

188 Lyon Sprague de Camp and Fletcher Pratt, *The Incomplete Enchanter* (New York: Henry Holt & Company, 1941).

platidlo“. Oživuje vaše svaly a nabíja vaše neuróny. Takmer každá metabolická reakcia závisí na ATP a teda na chemických vlastnostiach fosforu.

Ak prestanú fungovať zápalky, prestanete aj vy. Nedá sa zmeniť iba jedno z toho.

Povrchné pravidlá: „Zápalka sa zapáli po škrtnutí“ a „Ľudia potrebujú dýchať vzduch“ nie sú na prvý pohľad prepojené. Trvalo stáročia, kým sme toto spojenie objavili, a ešte aj teraz to vyzerá ako nejaký vzdialený fakt naučený v škole, dôležitý iba pre pár špecialistov. Je veľmi ľahké predstaviť si svet, v ktorom jedno povrchné pravidlo platí a druhé nie; potlačiť svoju vieru v jeden názor, ale nie v druhý. Ale to je *predstavivosť*, nie skutočnosť. Ak si rozdelíte mapu na štyri časti, aby sa vám lepšie skladovala, neznamená to, že aj územie sa rozdelilo na nesúvisiace časti. Naše mysle majú rôzne povrchné pravidlá odložené v rôznych priehradkách, ale to neodráža žiadne rozdelenie zákonov riadiacich Prírodu.

Môžeme v tejto lekcii zájsť ďalej. Fosfor odvodzuje svoje správanie od ešte hlbších zákonov, elektrodynamiky a chromodynamiky. „Fosfor“ je iba naše *slovo* pre elektróny a kvarky usporiadané určitým spôsobom. Nemôžete zmeniť chemické vlastnosti fosforu a nezmeniť pritom zákony riadiace elektróny a kvarky.

Keby ste vošli do sveta, v ktorom sa zápalky nedajú škrtnúť, prestali by ste existovať ako organizovaná hmota.

Realita je previazaná omnoho pevnejšie než sa ľuďom zdá.

*→

182. Všeobecný zákon

Antoine-Laurent de Lavoisier objavil, že dýchanie a oheň fungujú na rovnakom princípe. Bolo to jedno z najšokujúcejších zjednotení v histórii vedy, pretože spojilo pozemnú ríšu hmoty a posvätnú ríšu ohňa, ktoré ľudia dovtedy delili na samostatné magistériá.

Prvé veľké zjednodušenie urobil Isaac Newton, ktorý zjednotil dráhy planét s trajektóriou padajúceho jablka. Šok z tohto objavu bol ešte omnoho väčší než z Lavoisierovho. Nebolo to len tým, že sa Newton opovážil zjednotiť pozemskú ríšu nízkej hmoty s očividne odlišnou a posvätnou nebeskou ríšou, ktorá bola kedysi považovaná za príbytok bohov. Newtonov objav dal vzniknúť pojmu *všeobecného zákona*, zákona, ktorý je rovnaký všade a vždy, a ktorý má doslova *nula* výnimiek.

Ľudia žijú vo svete povrchných javov, a povrchné javy sú rozdelené na presakujúce kategórie s hromadou výnimiek. Tiger sa nespráva ako byvol. Väčšina byvolov má štyri nohy, ale možno tento konkrétny má iba tri. ň Prečo by si kto myslel, že môžu existovať zákony, ktoré platia všade? Očividne to tak nie je.

Jediný prípad, keď sa zdalo, že by sme *chceli* mať zákon, ktorý platí všade, je keď sa rozprávalo o morálnych zákonoch – pravidlách správania sa v tlupe. Niektorí členovia tlupy si možno chcú brať väčší než spravodlivý podiel z byvolieho mäsa – možno na to majú aj nejakú chytrú výhovorku – takže v prípade morálnych zákonov zrejme máme inštinkt univerzálnosti. Áno, pravidlo o delení mäsa rovnakým dielom sa týka aj *teba*, práve teraz, či sa ti to páči alebo nie. Ale aj tu sú výnimky. Keby sa – z nejakého divného dôvodu – omnoho silnejšia tlupa vyhrážala, že prebodne každého oštepom, pokiaľ Bob tento jedinýkrát nedostane dvakrát viac mäsa, dali by ste Bobovi dvakrát viac mäsa. Predstava pravidla, ktoré nemá doslova *žiadnu* výnimku, znie šialene prísne, ako výsledok myslenia uzavretej mysle fanatika, ktorý je natoľko posadnutý svojou jednou veľkou ideou, že nedokáže vnímať bohatstvo a zložitosť skutočného vesmíru.

Toto je tradičné obvinenie namierené voči vedcom – profesionálnym študentom bohatstva a zložitosti skutočného vesmíru. Pretože *keď sa na vesmír naozaj pozriete*, ukáže sa, že je podľa ľudských merítok šialene prísny v aplikovaní svojich pravidiel. Pokiaľ vieme, zatiaľ nenastalo *jedno jediné porušenie* zákona zachovania hybnosti, od počiatku času až dodnes.

→ http://lesswrong.com/lw/hq/universal_fire/

Niekedy – veľmi zriedkavo – pozorujeme domnelé porušenie našich *modelov* základných zákonov. Hoci naše vedecké modely vydržia generáciu alebo dve, nie sú stabilné v priebehu stáročí... ale nepredstavujte si, že toto robí samotný vesmír rozmarným. To by sme si miešali mapu s územím. Lebo keď sa prach usadí a stará teória je zvrhnutá, vysvitne, že vesmír sa *vždy* správal podľa nového zovšeobecnenia, ktoré sme objavili, ktoré je opäť absolútne všeobecné dokiaľ ľudské poznanie siaha. Keď sa objavilo, že newtonovská gravitácia je špeciálnym prípadom všeobecnej relativity, ukázalo sa, že všeobecná relativita riadila obeh Merkúra celé desaťročia predtým než o nej ľudia vedeli; a neskôr sa ukázalo, že všeobecná relativita riadila zrútenie hviezd celé miliardy rokov pred ľudstvom. To iba náš model sa mýlil – samotný Zákon bol vždy absolútne nemenný – aspoň to hovorí náš nový model.

Môžem vyjadriť 80%-nú istotu, že hranica rýchlosti svetla vydrží najbližších stotisíc rokov, ale to neznamená, že si myslím, že hranica rýchlosti svetla platí iba v 80% prípadov, s občasnými výnimkami. Výrok, ktorému prisudzujem pravdepodobnosť 80 % znie, že zákon o rýchlosti svetla je *absolútne neporušiteľný všade v priestore a čase*.

Jedným z dôvodov, prečo starovekí Gréci neobjavili vedu, je, že si neuvedomili, že na základe experimentov možno zovšeobecňovať. Gréckych filozofov zaujímali „normálne“ javy. Keby ste zostavili neprirodzený pokus, pravdepodobne by ste dostali „obludný“ výsledok, ktorý nemá žiaden vplyv na to, ako veci fungujú naozaj.

Takto teda ľudia snívajú, dokiaľ sa nenaučia lepšie; aké sú však tiché sny samotného vesmíru, ktoré sníval sám pre seba, dokiaľ sa mu neprisnili ľudia? Ak sa chcete naučiť rozmýšľať ako skutočnosť, potom je tu Tao:

Od samého začiatku

sa nestala jediná

nezvyčajná vec.

* →

183. Je skutočnosť škaredá?

Pozrime sa na tretie mocniny: { 1, 8, 28, 64, 125... }. Ich rozdiely { 7, 19, 37, 61... } môžu na prvý pohľad vyzerať, že nemajú zrejmý vzor, ale vezmime si rozdiely rozdielov { 12, 18, 24... } a dostaneme sa na úroveň jednoduchého vzťahu. Ak si vezmeme rozdiely rozdielov rozdielov { 6, 6... }, dostaneme sa na dokonale stabilnú úroveň, kde sa chaos rozpustil na poriadok.

Ale toto je úmyselne zvolený príklad. Možno „neusporiadanému skutočnému svetu“ chýba krása týchto matematických objektov? Možno by bolo vhodnejšie rozprávať o neurovede alebo sieťach expresie génov?

Abstraktná matematika, konštruovaná iba v predstavách, vzniká z jednoduchých základov – malá množina počiatočných axióm – a je uzavretý systém; podmienky, ktoré možno vyzerajú *neprirodzene* nápomocné elegancii.

Inými slovami: V čistej matematike sa nemusíte báť toho, že spoza kríka vyskočí tiger a zožerie Pascalov trojuholník.

Je teda skutočný svet škaredší než matematika?

Zvláštne, že sa to ľudia pýtajú. Chcem tým povedať, že táto otázka by dávala zmysel pred dva a pol tisícročím... Vtedy, keď Grécki filozofi debatovali o tom, z čoho by sa mohol tento „skutočný svet“ skladať, existovalo mnoho názorov. Hérakleitos povedal: „Všetko je oheň.“ Táles povedal: „Všetko je voda.“ Pytagoras povedal: „Všetko sú čísla.“

Skóre:

Hérakleitos	0
Táles	0

→ http://lesswrong.com/lw/hr//

Pod zložitými formami a tvarmi povrchného sveta existuje jednoduchá úroveň, presná a stabilná úroveň, ktorej zákony nazývame „fyzika“. Tento objav, toto Veľké Prekvapenie, už v našom bode ľudskej histórie bol urobený – ale nemali by sme zabúdať, že to bolo prekvapenie. Jedného dňa ľudia išli hľadať krásu za tým všetkým, bez záruky, že nejakú nájdu; a jedného dňa ju našli; a teraz je to známa vec, a berieme ju ako samozrejmosť.

Prečo teda nedokážeme predvídať polohu každého tigra v kríkoch rovnako ľahko ako predpovedáme túto šiestu tretiu mocninu?

Aj vo svetoch čírej matematiky vidím tri zdroje neistoty – dva zrejmé, jeden nie taký zrejmý.

Prvým zdrojom neistoty je, že aj tvor z čistej matematiky, žijúci vo svete z čistej matematiky, nemusí poznať túto matematiku. Ľudia chodili po Zemu dávno predtým ako Galileo/Newton/Einstein objavil zákon gravitácie, ktorý zabraňuje nášmu vyhodeniu do vesmíru. Stabilné základné zákony vás môžu riadiť bez toho, že by ste ich poznali. Neexistuje žiaden fyzikálny zákon hovoriaci, že fyzikálne zákony musia byť slovne vyjadrené, ako poznanie, v mozgoch, ktoré sa nimi riadia.

Zatiaľ ešte nemáme Teóriu všetkého. Naše súčasné najlepšie teórie sú vecou matematiky, ale nie sú navzájom dokonale prepojené. Najpravdepodobnejším vysvetlením je, že – ako sa ukázalo v minulosti – vidíme povrchné úkazy hlbšej matematiky. Zatiaľ je naším najlepším odhadom, že skutočnosť sa skladá z matematiky; zatiaľ však ešte celkom nevieme, ktorej matematiky.

Ale fyzici musia zostavovať obrovské urýchľovače častíc, aby rozlíšili medzi teóriami – aby prejavili svoju zostávajúci neistotu viditeľným spôsobom. Že fyzici musia zájsť takto ďaleko, aby si neboli istí, naznačuje, že toto nie je zdrojom našej neistoty ohľadom cien na burze.

Druhým zrejmým zdrojom neistoty je, že aj keď poznáte všetky súvisiace fyzikálne zákony, nemusíte mať dosť výpočtovej sily na ich extrapoláciu. Poznáme každý základný fyzikálny zákon súvisiaci s tým, ako sa reťaz aminokyselín skladá na bielkovinu. Ale stále nedokážene predpovedať tvar bielkoviny podľa aminokyselín. Nejaká mrňavá 5-nanometrová molekula, ktorá sa zloží za mikrosekundu, je *príliš veľa informácie* na to, aby ju dnešné počítače zvládli (a to nehovorím o tigroch a cenách na burze). Naše najmodernejšie úsilie v skladaní bielkovím využíva chytré aproximácie, nie základnú Schrödingerovu rovnicu. Keď dôjde na opis 5-nanometrového predmetu pomocou *naozaj* základnej fyziky, kvarkov – no, ani sa neunúvajte to skúšať.

Musíme používať nástroje ako rentgenovú kryštalografiu a NMR, aby sme odhalili tvary bielkovín, ktoré sú plne určené známymi fyzikálnymi zákonmi a známou postupnosťou DNA. Nie sme logicky vševedúci; nedokážeme vidieť dôsledky našich myšlienok; nevieme, čomu veríme.

Tretí zdroj neistoty je ten najťažší na pochopenie, a Nick Bostrom o tom napísal knihu. Predstavte si, že postupnosť { 1, 8, 27, 64, 125... } existuje; predstavte si, že je to fakt. A predstavte si, že navrchu každej tretej mocniny je malý človiečik – jeden človiečik na každej tretej mocnine – a predstavte si, že aj toto je fakt.

Ak stojíte mimo a pozeráte sa z globálneho hľadiska – pozeráte sa z výšky na túto postupnosť tretích mocnín a malých ľudkov usadených navrchu – potom tieto dva fakty hovoria všetko, čo sa dá vedieť o tejto postupnosti a o týchto ľudkoch.

Ale ak ste jeden z tých malých ľudkov sediaci na tretej mocnine, a ak viete tieto dva fakty, stále je tu tretia časť informácie, ktorú potrebujete, aby ste vedeli robiť predpovede: „Na ktorej tretej mocnine stojím *ja*?“

Očakávate, že sa ukáže, že stojíte na tretej mocnine; neočakávate, že sa ukáže, že stojíte na čísle 7. Vaše očakávania sú jednoznačne ohraničené vašimi znalosťami základnej fyziky; vaše názory sú falzifikovateľné. Ale aj tak sa musíte pozrieť pod seba a zistiť, či stojíte na 1728 alebo na 5177717. Ak viete rýchlo spamäti počítať a uvidíte, že prvé dve číslice zo štvorcifernej tretej mocniny sú 17__, stačí vám to na uhádnutie, že zvyšné dve číslice sú 2 a 8. V opačnom prípade sa musíte pozrieť, aby ste objavili aj 2 a 8.

Aby ste zistili, ako by mala vyzerať nočná obloha, nestačí vám poznať fyzikálne zákony. Nestačí vám ani mať logickú vševedúcnosť o ich dôsledkoch. Musíte vedieť, *kde* vo vesmíre sa nachádzate. Musíte vedieť, že sa pozeráte na nočnú oblohu *zo Zeme*. Potrebujete informácie nielen na nájdenie Zeme vo *viditeľnom* vesmíre, ale v celom vesmíre, vrátane tých častí, ktoré naše teleskopy nevidia, pretože sú príliš vzdialené, a odlišných inflačných vesmírov, a alternatívnych Everettových vetiev.

Je dobrý tip, že „neistota ohľadom počiatočných podmienok na hranici“ je v skutočnosti indexická neistota. Ale ak nie, je to empirická neistota, neistota o tom, aký *je* vesmír z globálneho pohľadu, čo ju dáva do rovnakej kategórie ako je neistota ohľadom fyzikálnych zákonov.

Kdekoľvek je náš najlepší tip, že „skutočný svet“ má *nezachrániteľne* neusporiadanú zložku, je to kvôli tomuto druhému a tretiemu zdroju neistoty – logická neistota a indexická neistota.

Neznalosť základných zákonov vám nepovie, že neusporiadane vyzerajúci vzor je naozaj neusporiadaný. Môže sa stať, že ste zatiaľ ešte len neobjavili poriadok.

Ale keď ide o neusporiadané siete expresie génov, *už sme našli* tú skrytú krásu – stabilnú úroveň fyziky za tým. *Pretože* sme už našli hlavný poriadok, môžeme odhadovať, že už nenájdeme žiadne *ďalšie* tajné vzorce, ktoré by urobili biológiu rovnako ľahkou ako postupnosť tretích mocnín. Poznáme už pravidlá hry, vieme, že tá hra je ťažká. Nemáme dosť výpočtovej sily, aby sme urobili proteínovú chémiu podľa fyziky (druhý zdroj neistoty) a evolučné dráhy mohli na rôznych planétach ísť rôznym smerom (tretí zdroj neistoty). Nové objavy v základnej fyzike nám tu nepomôžu.

Keby ste boli starým Grékom hľadiacim na surové údaje z biologického pokusu, dávalo by zmysel, keby ste hľadali skrytú štruktúru s pytagorovskou eleganciou, všetky bielkoviny zoradené v dokonalom dvadsaťstene. Lenže v biológii už vieme, kde je táto pytagorovská elegancia, a vieme, že je príliš hlboko dole než aby nám pomohla prekonať našu indexickú a logickú neistotu.

Podobne si môžeme byť istí, že nikto nikdy nebude vedieť predpovedať výsledky istých kvantových pokusov, pretože naša základná teória nám pomerne definitívne hovorí, že naše rôzne verzie uvidia rôzne výsledky. Ak vám znalosť základných zákonov hovorí, že existuje postupnosť tretích mocnín a že navrchu každej tretej mocniny stojí malý človiečik, a že títo malí ľudkovia sú rovnakí až na to, že sú na rôznych tretích mocninách, a že vy ste jeden z týchto ľudkov, potom *viete*, že nemáte žiaden iný spôsob ako zistiť, na ktorej tretej mocnine ste, okrem toho, že sa pozriete.

Najlepšie súčasné vedomosti hovoria, že „skutočný svet“ je dokonale pravidelný, deterministický, a *veľmi veľký* matematický objekt, ktorý je veľmi drahé simulovať. Takže „skutočný svet“ je menej ako predpovedanie ďalšej tretej mocniny v postupnosti tretích mocnín, a viac ako vedieť, že navrchu týchto tretích mocnín stojí veľa malých ľudkov, ale nevedieť, ktorý z nich ste *vy osobne*, a zároveň nebyť veľmi dobrý v počítaní spamäti. Naše znalosti pravidiel obmedzujú naše očakávania, celkom dosť, ale nie dokonale.

A neznie vari *toto* ako skutočný život?

Neurčitosť však existuje na mape, nie v území. Ak nevieme o nejakom jave, je to fakt o našom stave mysle, nie fakt o samotnom jave. Empirická neistota, logická neistota, a indexická neistota sú iba názvy pre náš vlastný zmätok. Najlepší súčasný tip je, že svet je matematika a matematika je dokonale pravidelná. Neusporiadanosť je iba v očiach pozorovateľa.

* →

184. Krásna pravdepodobnosť

Mali by sme očakávať, že rozumnosť bude, *na istej úrovni*, jednoduchá? Mali by sme hľadať a dúfať v *základnú* krásu za umením veriť a vyberať si?

Dovoľte mi požičať si na úvod tejto témy sťažnosť neskorého veľkého bayesiánskeho majstra, E. T. Jaynesa:[189]

→ http://lesswrong.com/lw/ms/is_reality_ugly/

189 Edwin T. Jaynes, „Probability Theory as Logic,“ [Teória pravdepodobnosti ako logika] in *Maximum Entropy and Bayesian Methods*, [Maximálna entropia a bayesovské metódy] ed. Paul F. Fougère (Springer Netherlands, 1990).

> „Dvaja lekárski výskumníci používajú tú istú terapiu nezávisle, v rôznych nemocniciach. Ani jeden z nich by sa neznížil k falšovaniu údajov, ale jeden z nich sa vopred rozhodol, že vzhľadom na obmedzené zdroje prestane liečiť po N = 100 pacientoch, bez ohľadu na to, koľko prípadov vyliečenia bude dovtedy vidieť. Druhý stavil svoju povesť na účinnosť liečby, a rozhodol sa, že sa nezastaví, dokiaľ nebude mať údaje naznačujúce, že miera vyliečenia je jednoznačne vyššia než 60 %, bez ohľadu na to, koľkých pacientov bude musieť liečiť. V skutočnosti obaja skončili s úplne rovnakými údajmi: n = 100 [pacientov], r = 70 [vyliečení]. Mali by sme teda z ich pokusov vyvodiť rôzne výsledky?“ (Predpokladajme, že obe kontrolné skupiny tiež mali rovnaké výsledky.)

Cyan nás odkazuje na 37. kapitolu MacKayovej vynikajúcej učebnice štatistiky, voľne dostupnej na internete, kde je dôkladnejšie vysvetlenie tohto problému.[190]

Podľa klasickej štatistickej procedúry – ktorá sa, pokiaľ viem, vyučuje dodnes – títo dvaja výskumníci urobili rôzne pokusy s rôznymi koncovými podmienkami. Tieto dva pokusy *mohli* skončiť s rôznymi údajmi, a preto predstavujú rôzne testy hypotézy, a vyžadujú si rôzne štatistické analýzy. Je celkom možné, že prvý pokus bude „štatisticky významný“ a druhý nie.

Či vás toto znepokojuje alebo nie, vypovedá veľa o vašom postoji k teórii pravdepodobnosti a vlastne k rozumnosti samotnej.

Nebayesovský štatistik by mohol pokrčiť plecami a povedať: „Nuž, nie všetky štatistické nástroje majú rovnaké silné a slabé stránky, chápete – kladivo nie je skrutkovač – a ak použijete rôzne štatistické nástroje, môžete dostať rôzne výsledky, rovnako ako keď použijete tie isté údaje na výpočet lineárnej regresie alebo na výcvik regularizovanej neurónovej siete. Musíte použiť správny nástroj pre danú situáciu. Život je neusporiadaný...“

A tu je bayesovská odpoveď: „*Prepáčte?* Dopad indícií pevne danej pokusnej metódy produkujúcej tie isté údaje závisí od súkromných myšlienok výskumníka? A vy máte tú drzosť označovať *nás* za ‚príliš subjektívnych‘?“

Ak je príroda v jednom stave, podmienená pravdepodobnosť, že údaje vyjdú tak, ako sme videli, bude jedna vec. Ak je príroda v inom stave, podmienená pravdepodobnosť, že údaje vyjdú takto, bude niečo iné. Ale podmienená pravdepodobnosť, že daný stav prírordy vytvorí údaje, ktoré sme videli, nemá nič spoločné s výskumníkovými súkromnými úmyslami. Takže bez ohľadu na naše hypotézy o prírode, podmienená pravdepodobnosť je rovnaká, a dopad indícií je rovnaký, a naše výsledné názory by mali byť rovnaké, pri oboch pokusoch. Aspoň jedna z dvoch klasických metód musela zahodiť dôležité informácie - alebo jednoducho nesprávne počítať - ak tieto dve metódy došli k rôznym záverom.

Pradávna vojna medzi bayesovcami a zlorečenými frekventistami sa ťahá celé desaťročia, a nebudem sa ani pokúšať zreprodukovať túto dávnu históriu v tomto článku.

Ale jeden z ústredných konfliktov je, že bayesovci očakávajú, že teória pravdepodobnosti bude... aké slovo to hľadám? „Pekná?“ „Čistá?“ „Vnútorne konzistentná?“

Ako hovorí Jaynes, vety bayesovskej pravdepodobnosti sú jednoducho toto, *vety* koherentného systému dôkazov. Bez ohľadu na to, aké odvodenia používate, v akom poradí, výsledky bayesovskej pravdepodobnosti by mali byť vždy konzistentné – každá veta kompatibilná s každou inou vetou.

Ak chcete vedieť súčet 10 + 10, môžete si ho predefinovať ako (2 * 5) + (7 + 3) alebo ako (2 * (4 + 6)) alebo použiť ľubovoľné iné *legálne* triky, ale výsledok musí vždy vyjsť rovnaký, v tomto prípade 20. Ak niečo vyjde jedným spôsobom ako 20 a druhým ako 19, potom môžete uzavrieť, že ste prinajmenšom v jednom z týchto dvoch prípadov urobili niečo nelegálne. (V aritmetike je tou nelegálnou operáciou zvyčajne delenie nulou; v teórii pravdepodobnosti je to zvyčajne nekonečno, ktoré nebolo použité ako limita konečného procesu.)

Ak dostanete výsledok 19 = 20, poriadne hľadajte, kde ste urobili chybu, pretože je nepravdepodobné, že ste práve vyhodili do vzduchu samotnú aritmetiku. Keby sa niekomu niekedy

190 David J. C. MacKay, *Information Theory, Inference, and Learning Algorithms* [Teória informácií, odvodzovanie, a učiace sa algoritmy] (New York: Cambridge University Press, 2003)

naozaj podarilo odvodiť *skutočný* rozpor z bayesovskej teórie pravdepodobnosti – napríklad dva rôzne empirické dopady z tej istej pokusnej metódy dávajúcej tie isté výsledky – celá táto stavba by vyletela do vzduchu. Spolu s teóriou množín, pretože som si dosť istý, že ZF poskytuje model pre teóriu pravdepodobnosti.

Matematika! To je to slovo, ktoré som hľadal. Bayesovci očakávajú, že teória pravdepodobnosti bude *matematika*. Preto sa zaujímame o Coxovu vetu a jej mnohé rozšírenia ukazujúce, že každá reprezentácia neistoty, ktorá dodržiava určité obmedzenia, sa musí mapovať na teóriu pravdepodobnosti. Koherentná matematika je skvelá, ale jednoznačná matematika je ešte lepšia.

A predsa... *mala by* rozumnosť byť matematikou? Záver, že by pravdepodobnosť mala byť pekná, rozhodne nie je samozrejmý. Skutočný svet je neusporiadaný – nemali by sme na jeho zvládanie používať neusporiadané myslenie? Možno nebayesovskí štatistici s ich rozľahlou zbierkou ad-hoc metód a ad-hoc zdôvodnení sú jasne kompetentnejší, pretože majú jasne väčšiu zbierku nástrojov. Je pekné, keď sú problémy čisté, ale zvyčajne nie sú, a s tým musíme žiť.

Napokon, je všeobecne známe, že pri mnohých problémoch nemôžete použiť bayesovské metódy, pretože bayesovské výpočty sú výpočtovo nezvládnuteľné. Prečo teda nenechať kvitnúť mnoho kvetín? Prečo nemať vo svojom kufríku viac nástrojov?

Toto je základný rozdiel v spôsobe myslenia. Štatistici starej školy mysleli v pojmoch *nástrojov*, trikov, ktoré sa používajú na konkrétne problémy. Bayesovci – prinajmenšom tento jeden bayesovec, hoci si nemyslím, že hovorím iba za seba – rozmýšľajú o *zákonoch*.

Hľadanie zákonov nie je to isté ako hľadanie zvlášť šikovných a pekných nástrojov. Druhý termodynamický zákon nie je šikovná a pekná chladnička.

Carnotov cyklus je ideálny stroj – vlastne *jediný* ideálny stroj. Žiaden stroj poháňaný dvoma zásobníkmi tepla nemôže byť efektívnejší ako Carnotov stroj. V dôsledku toho sú všetky termodynamicky reverzibilné stroje fungujúce medzi rovnakými zásobníkmi tepla rovnako účinné.

Ale samozrejme nemôžete použiť Carnotov stroj na poháňanie skutočného auta. Stroj skutočného auta sa na Carnotov stroj podobá natoľko, ako sa pneumatiky auta podobajú na dokonalé otáčajúce sa valce.

Je teda jasné, že Carnotov stroj je nepoužiteľný *nástroj* na budovanie auta v skutočnom svete. Druhý termodynamický zákon sa tu samozrejme nedá použiť. Je ťažké vytvoriť stroj, ktorý sa ním v skutočnom svete riadi. Radšej ignorujme termodynamiku – použime čokoľvek, čo funguje.

Takýto typ zmätenia podľa mňa panuje nad tými, ktorí sa stále držia starej cesty.

Nie, nemôžete vždy urobiť presný bayesovský výpočet daného problému. Niekedy musíte hľadať približné riešenie; vlastne dosť často. To znamená, že teória pravdepodobnosti prestala platiť, rovnako ako vaša neschopnosť spočítať aerodynamiku lietadla 747 atóm po atóme neznamená, že sa lietadlo 747 neskladá z atómov. Nech použijete hocijaký približný výpočet, funguje do tej miery, do akej sa približuje ideálnemu bayesovskému výpočtu – a zlyháva do tej miery, do akej sa odchyľuje.

Dôkazy koherencie a jedinečnosti bayesiánstva fungujú oboma smermi. Rovnako ako sa každý výpočet, ktorý dodržiava Coxove axiómy koherencie (alebo niektoré z mnohých preformulovaní a zovšeobecnení), musí mapovať na pravdepodobnosti, takisto všetko, čo nie je bayesovské, musí zlyhať v niektorom z testov koherentnosti. Toto vás potom vystavuje trestom ako Dutch-booking (prijímanie kombinácií stávok, pri ktorých určite strácate, alebo odmietanie kombinácií stávok, pri ktorých by ste určite zarobili).

Možno nedokážete vypočítať optimálnu odpoveď. Ale keď použijete hocijaký približný výpočet, jeho zlyhania aj úspechy sa budú dať *vysvetliť* v pojmoch bayesovskej teórie pravdepodobnosti. Možno toto vysvetlenie nepoznáte; to neznamená, že to vysvetlenie neexistuje.

Chcete teda namiesto robenia bayesovských aktualizácií použiť lineárnu regresiu? Pozrite sa na štruktúru za lineárnou regresiou a uvidíte, že zodpovedá výberu najlepšieho bodového odhadu pri danej gausssovskej funkcii podmienenej pravdepodobnosti a uniformnej pôvodnej pravdepodobnosti parametrov.

Chcete použiť regularizovanú lineárnu regresiu, lebo tá v praxi lepšie funguje? Nuž, toto zodpovedá (bayesiánskymi slovami) gaussovskej pôvodnej pravdepodobnosti nad váhami.

Niekedy nemôžete používať bayesovské metódy *doslova*; vlastne dosť často. Ale keď *môžete* použiť presné bayesovské výpočty, ktoré využijú každý kúsok dostupného poznania, potom ste hotoví. Nikdy nenájdete štatistickú metódu, ktorá dá *lepší* výsledok. Možno nájdete lacný približný výpočet, ktorý takmer vždy funguje vynikajúco, a je lacnejší, ale nebude presnejší. Nie, pokiaľ tá druhá metóda nepoužíva nejaké poznanie, možno vo forme nevysloveného pôvodného predpokladu, ktoré ste nepustili do svojho bayesovského výpočtu; a potom, keď túto prvotnú informáciu vložíte do bayesovského výpočtu, bayesovský výpočet bude opäť rovnaký alebo lepší.

Keď používate klasické ad-hoc štatistické nástroje s ad-hoc (hoci často veľmi zaujímavým) zdôvodnením, nikdy neviete, či niekto zajtra nepríde s ešte chytrejším nástrojom. Ale keď *môžete* priamo použiť výpočet, ktorý odráža bayesovské zákony, ste *hotoví* – ako keby sa vám podarilo vložiť do auta Carnotov tepelný stroj. Je to, ako sa hovorí, „bayesovsky optimálne“.

Pripadá mi, akoby sa priaznivci nástrojov pozerali na postupnosť tretích mocnín {1, 8, 27, 64, 125...} a ukazovali na ich rozdiely {7, 19, 37, 61...} a hovorili: „Vidíte, život nie je vždy taký jednoduchý – musíte sa prispôsobiť okolnostiam.“ A bayesovci ukazujú na rozdiely rozdielov rozdielov, tú základnú stabilnú úroveň {6, 6, 6, 6, 6...}. A kritici hovoria: „O čom to do čerta hovoríte? Je to 7, 19, 37, nie 6, 6, 6. Príliš zjednodušujete tento neusporiadaný problém; ste príliš pripútaní k jednoduchosti.“

Nemusí to byť nevyhnutne jednoduché na *povrchnej* úrovni. Musíte sa ponoriť hlbšie, aby ste našli pevnú pôdu.

Myslite na zákony, nie nástroje. To, že musíte robiť približné výpočty nejakého zákona, nemení tento zákon. Lietadlá sú stále atómy, neriadia sa špeciálnymi výnimkami v aerodynamických zákonoch prírody. Približnosť existuje na mape, nie v území. Môžete poznať druhý termodynamický zákon, a predsa pracovať ako inžinier a stavať nedokonalé motory do áut. Druhý zákon neprestane platiť; vaša znalosť tohto zákona a Carnotových cyklov vám pomôže dostať sa tak blízko k ideálnej účinnosti, ako sa dá.

Nie sme očarení bayesovskými metódami iba preto, že sú krásne. Krása je vedľajší účinok. Bayesovské *vety* sú elegantné, koherentné, optimálne, a dokázateľne jedinečné, pretože sú to *zákony*.

*→

185. Mimo laboratória

„Mimo laboratória nie sú vedci o nič múdrejší než hocikto iný.“ Niekedy toto príslovie citujú vedci, skromne, smutne, aby si pripomenuli svoju vlastnú omylnosť. Niekedy sa toto príslovie cituje z menej chvályhodných dôvodov, aby sa znevážila nežiadúca odborná rada. Je toto príslovie pravdivé? Doslovne vzaté, asi nie. Znie príliš pesimisticky povedať, že vedci doslova nie sú *o nič* múdrejší ako priemer, že korelácia je doslova *nulová*.

Toto príslovie však do istej miery vyzerá pravdivo a myslím si, že by sme touto skutočnosťou mali byť veľmi znepokojení. Nemali by sme vzdychnúť a smutne zvesiť hlavu. Skôr by sme sa mali prebudiť a vzpriamiť. Prečo? Nuž, predstavte si, že bačovského učňa namáhavo naučíte počítať ovce, ako vychádzajú z ohrady a vchádzajú do nej. Bača takto vie, kedy všetky ovce odišli a kedy sa všetky ovce vrátili. Potom dáte bačovi pár jabĺk a poviete: „Koľko je to jabĺk?“ Ale bača na vás neprítomne hľadí, pretože ho nenaučili počítať jablká – iba ovce. Asi by ste baču podozrievali, že počítaniu veľmi nerozumie.

Teraz si predstavte, že zistíme, že doktor ekonómie si každý týždeň kupuje žreb lotérie. Musíme si položiť otázku: Naozaj tento človek *rozumie* očakávanému úžitku, do hĺbky? Alebo ho len naučili vykonávať určité matematické triky?

Pripomeňme si, ako Richard Feynman opisoval neúspešný program vyučovania fyziky:

→ http://lesswrong.com/lw/mt/beautiful_probability/

> „Títo študenti sa všetko naučili naspamäť, nevedeli však, čo tie veci znamenajú. Keď počuli ‚svetlo, ktoré sa odrazí od média s indexom,‘ nevedeli, že sa tým myslí materiál *ako napríklad voda*. Nevedeli, že ‚smer svetla‘ je smer, z ktorého niečo *vidíte*, keď sa na to pozeráte, a tak ďalej. Všetko bolo naučené úplne naspamäť, nič z toho však nebolo preložené do zmysluplných slov. Takže keď sa opýtam: ‚Čo je Brewsterov uhol?‘, pristupujem k počítaču so správnymi kľúčovými slovami. Ale keď poviem: ‚Pozrite na vodu,‘ nič sa nestane – pod heslom ‚Pozrite na vodu‘ nemajú nič.“

Predstavte si, že máme napohľad schopného vedca, ktorý vie ako navrhnúť pokus na N subjektoch; týchto N subjektov dostane náhodne pridelenú liečbu; porota nepoznajúca rozdelenie vyhodnotí výsledky subjektov; potom zbehneme výsledky na počítači a pozrieme, či sú významné na úrovni spoľahlivosti 0,05. Toto nie je iba tradičný rituál. Toto nie je otázka svojvoľnej etikety, ako používanie správnej vidličky na šalát. Je to tradičný rituál na *experimentálne testovanie hypotéz*. Prečo by ste mali experimentálne testovať svoje hypotézy? Pretože viete, že to od vás bude časopis vyžadovať pred uverejnením vášho článku? Pretože tak vás to naučili robiť na vysokej? Pretože všetci jednohlasne hovoria, že je dôležité urobiť experiment, a budú sa na vás čudne pozerať, ak poviete niečo iné?

Nie: pretože, aby ste vytvorili mapu územia, musíte ísť von a pozrieť sa na to územie. Nie je možné nakresliť presnú mapu mesta, kým sedíte vo svojej obývačke so zatvorenými očami, myslíte príjemné myšlienky o tom, ako by si želali, aby mesto vyzeralo. Musíte ísť von, prechádzať cez mesto, a značiť na papier čiary zodpovedajúce tomu, čo vidíte. V maličkej miere sa to deje zakaždým, keď sa pozriete na svoje topánky, či máte šnúrky zaviazané. Fotóny prídu zo Slnka, odrazia sa od vašich šnúrok, dopadnú na vašu sietnicu, preložia sa na frekvencie neurónových impulzov a zrekonštruujú sa vašou zrakovou kôrou na aktivačný vzor, ktorý silno koreluje so súčasným stavom vašich šnúrok na topánkach. Aby ste získali nové informácie o území, musíte interagovať s týmto územím. Musí existovať nejaký skutočný, fyzikálny proces, ktorým váš mozog skončí korelovaný so stavom prostredia. Procesy rozmýšľania nie sú čarovné; môžete kauzálne popísať, ako fungujú. A z tohto všetkého vyplýva, že ak chcete veci zistiť, musíte sa ísť pozrieť.

Čo si máme teraz pomyslieť o vedcovi, ktorí v laboratóriu vyzerá schopný, ale mimo laboratória verí v svet duchov? Pýtame sa, prečo, a vedec povie niečo v duchu: „Nuž, nikto naozaj nevie, a pripúšťam, že nemám žiadne indície – je to náboženské presvedčenie, nedá sa vyvrátiť takým ani onakým pozorovaním.“ Musel by som dôjsť k záveru, že tento človek *doslova nevie, prečo sa treba na veci pozerať*. Mohli ho naučiť určitý rituál experimentovania, ale nerozumie jeho *dôvodom* – že pri tvorbe mapy územia sa naň musíte pozrieť – že aby ste získali informáciu o prostredí, musíte podstúpiť kauzálny proces, ktorým interagujete s prostredím a skončíte s ním korelovaný. To sa rovnako týka návrhu dvojito zaslepeného experimentu, ktorý zbiera informácie o účinnosti novej liečebnej pomôcky, ako vašich očí, ktoré zbierajú informácie o vašich šnúrkach na topánkach.

Možno náš duchovný vedec povie: „Ale toto nie je vec na experimentovanie. Duchovia ku mne hovoria v mojom srdci.“ Nuž, ak naozaj predpokladáme, že duchovia komunikujú, ľubovoľným spôsobom, je to kauzálna interakcia a tá sa počíta ako pozorovanie. Stále platí teória pravdepodobnosti. Ak tvrdíte, že nejaká osobná skúsenosť s „hlasmi duchov“ je indíciou, že duchovia naozaj existujú, musíte tvrdiť, že existuje pomer pravdepodobnosti v prospech duchov spôsobujúcich „hlasy duchov“, v porovnaní s inými vysvetleniami „hlasov duchov“, dostatočný na prekonanie prvotnej nepravdepodobnosti zložitého vysvetlenia skladajúceho sa z mnohých častí. Neuvedomiť si, že „duchovia ku mne hovoria v mojom srdci“ je príklad „kauzálnej interakcie“ je analogické k tomu, keď si študent fyziky neuvedomuje, že „médium s indexom“ znamená materiál ako je voda.

Možno je jednoduché nechať sa oklamať skutočnosťou, že ľudia nosiaci laboratórne plášte používajú slovné spojenie „kauzálna interakcia“ a že ľudia nosiaci krikľavé šperky používajú slovné spojenie „hlasy duchov“. Ako všetci vieme, diskutujúci nosiaci odlišné oblečenie vymedzujú nezávislé oblasti existencie - „oddelené magistériá“, použijúc nesmrteľné nevydarené slovné spojenie Stephena J.

Goulda. V skutočnosti, „kauzálna interakcia“ je len ozdobný spôsob ako povedať: „Niečo spôsobuje, že sa niečo iné stane,“ a teória pravdepodobnosti kašle na to, aké oblečenie nosíte.

V modernej spoločnosti prevláda názor, že duchovné záležitosti nemožno vyriešiť logikou ani pozorovaním, a preto môžete mať náboženské presvedčenie, aké sa vám páči. Ak sa na toto nejaký vedec nechá nachytať a rozhodne sa žiť svoj mimolaboratórny život podľa toho, pre mňa to potom znamená, že chápe experimentálny princíp len ako *spoločenskú zvyklosť*. Vie, kedy sa od neho *očakáva*, že bude robiť experimenty a testovať štatistickú významnosť výsledkov. Ale dajte ho do situácie, kde je *spoločenskou zvyklosťou* vymýšľať si podivné názory bez pozerania sa, a on bude rovnako ochotne robiť aj toto.

Bačovskému učňovi povedia, že ak „sedem“ oviec vyjde von, a „osem“ oviec vyjde von, potom by „pätnásť“ oviec malo prísť naspäť. Prečo „pätnásť“ a nie „štrnásť“ alebo „tri“? Pretože inak dnes nedostaneš večeru, takže preto! Toto je istý druh odborného výcviku, a funguje podľa módy – ale ak je spoločenská zvyklosť jediným dôvodom, prečo sa sedem oviec plus osem oviec rovná pätnásť oviec, možno sa sedem jabĺk plus osem jabĺk rovná tri jablká. Veď kto hovorí, že pravidlá pre jablká nemôžu byť iné?

Ale ak viete, *prečo* tieto pravidlá fungujú, vidíte, že sčítanie je rovnaké pre ovce aj pre jablká. Isaac Newton je právom uctievaný, nie pre jeho zastaranú teóriu gravitácie, ale pre objav, že – úžasné, prekvapujúce – nebeské telesá v slávnych nebesiach poslúchajú tie isté zákony ako padajúce jablká. V makroskopickom svete – každodennom životnom prostredí našich predkov – rôzne stromy dávajú rôzne ovocie, rôzne zvyky platia pre rôznych ľudí v rôznych časoch. Skutočne zjednotený vesmír, s nemennými vesmírnymi zákonmi, je pre človeka vysoko neintuitívna predstava! Iba vedci tomu naozaj veria, aj keď niektoré náboženstvá toho narozprávajú o „jednote všetkých vecí“.

Ako povedal Richard Feynman:

> „Ak sa dostatočne zblízka pozrieme na pohár, uvidíme celý vesmír. Tu sú tie veci z fyziky: prúdiaca tekutina, ktorá sa vyparuje podľa vetra a počasia, odrazy v skle, a naša predstavivosť pridáva atómy. Sklo je destilátom skál Zeme a v jeho zložení vidíme tajomstvo veku vesmíru a vývoja hviezd. Aké zvláštne zoskupenie chemických látok je vo víne? Ako sa tam dostali? Sú tam kvasinky, enzýmy, substráty a produkty. Tam vo víne nájdeme veľké zovšeobecnenie: všetok život je kvasenie. Kto objaví chémiu vína, ten musí objaviť, tak ako Louis Pasteur, príčiny mnohých chorôb. Aká svieža je jeho chuť, vtláčajúca svoju existenciu do vedomia, ktoré ho pozoruje! Ak naše malé mysle kvôli pohodlnosti rozdeľujú tento pohár vína, tento vesmír, na časti – fyzika, biológia, geológia, astronómia, psychológia, a tak ďalej – pamätajme, že Príroda o tomto nevie! Poskladajme to teda všetko späť dokopy, aby sme nakoniec nezabudli, na čo to je. Doprajme si ešte jedno záverečné potešenie: vypime ho a zabudnime na všetko!“

Pár náboženstiev, najmä tých vymyslených alebo zrekonštruovaných po Isaacovi Newtonovi, môže vyznávať, že „všetko súvisí so všetkým“. (Keďže existuje triviálny izomorfizmus medzi grafmi a ich komplementmi, táto hlboká múdrosť vyjadruje rovnako užitočnú informáciu ako graf bez hrán.) Ale keď príde na skutočnú podstatu náboženstva, proroci a kňazi nasledujú prastarú ľudskú tradíciu vymyslieť si všetko za chodu. A tak si vymyslia jedno pravidlo pre ženy do dvanásť rokov, iné pravidlo pre mužov nad tridsať; jedno pravidlo pre sobotu a iné pravidlo na pracovný deň; jedno pravidlo pre vedu a iné pravidlo pre čarodejníctvo...

Skutočnosť, ako sme sa na vlastné prekvapenie naučili, *nie je* zbierkou oddelených magistérií, ale jeden zjednotený proces riadený na nízkej úrovni jednoduchými matematickými pravidlami. Rôzne budovy na pôde univerzity nepatria do rôznych vesmírov, aj keď to tak občas môže vyzerať. Vesmír nie je rozdelený na myseľ a hmotu, alebo život a neživot; atómy v našich hlavách bez problémov interagujú s atómami okolitého vzduchu. Ani Bayesova veta sa nelíši z jedného miesta na druhé.

Ak mimo oblasti svojej vedeckej špecializácie je nejaký konkrétny vedec rovnako náchylný k podivným myšlienkam ako každý iný, potom pravdepodobne nikdy nepochopil, *prečo* fungujú pravidlá vedy. Možno dokáže odpapagájovať niečo o Popperovskej falzifikácii, ale nerozumie jej do hĺbky, na

algebraickej úrovni teórie pravdepodobnosti, na kauzálnej úrovni poznania ako stroja. Naučili ho správať sa v laboratóriu určitým spôsobom, nemá však *rád*, keď ho obmedzujú indície; keď ide domov, vyzlečie si laboratórny plášť a uvoľní sa nejakým pohodlným nezmyslom. A áno, potom sa čudujem, či môžem dôverovať názorom tohto vedca aj v jeho vlastnej oblasti – najmä ak ide o nejakú kontroverznú tému, otvorenú otázku, čokoľvek čo nie už dávno zaklincované hromadou indícií a spoločenskou zvyklosťou.

Možno toto príslovie *dokážeme* prekonať – byť rozumní vo svojich osobných životoch, nielen vo svojich profesionálnych životoch. Nemalo byť nás zastaviť púhe príslovie: „Vtipný výrok nič nedokazuje," povedal Voltaire. Možno máme na viac, ak naštudujeme dosť teórie pravdepodobnosti, aby sme vedeli, *prečo* tie pravidlá fungujú, a dosť experimentálnej psychológie, aby sme videli, ako sa uplatňujú v prípadoch zo skutočného života – ak sa naučíme pozerať na vodu. Takejto ambícii chýba pohodlná skromnosť môcť vyznať, že mimo vašej špecializácie nie ste o nič lepší ako hocikto iný. Ale ak sa naše teórie o rozumnosti nedajú zovšeobecniť na každodenný život, potom niečo robíme nesprávne. V laboratóriu a mimo laboratória nie sú odlišné vesmíry.

*→

186. Druhý zákon termodynamiky a stroje na poznanie

Prvý zákon termodynamiky, viac známy ako zákon zachovania energie, hovorí, že nemôžete vytvoriť energiu z ničoho: zakazuje stroje na večný pohyb prvého typu, ktoré bežia a bežia donekonečna bez spotreby paliva alebo ľubovoľnej inej suroviny. Podľa nášho moderného chápania fyziky sa energia zachováva v každej *jednotlivej* interakcii častíc. Podľa matematickej indukcie vidíme, že nezáleží na tom, aké veľké zhromaždenie častíc máme, nedokáže tvoriť energiu z ničoho – nie bez porušenia toho, čo dnes považujeme za zákony fyziky.

To je dôvod, prečo americký patentový úrad súhrnne zamietne váš úžasne chytrý návrh sústavy kolies a ozubených kolies, ktoré navíjajú jednu pružinu kým sa druhá rozvíja, a budú to robiť donekonečna, podľa vašich výpočtov. Existuje *úplne všeobecný* dôkaz, že aspoň jedno koliesko musí porušovať zákony fyziky (náš štandardný model), aby toto bolo možné. Takže ak neviete vysvetliť, ako *jedno* koliesko porušuje zákony fyziky, nepomôže vám ani celá *sústava* kolies.

Podobný argument sa týka „nereaktívneho pohonu", systému pohonu, ktorý porušuje zákon zachovania hybnosti. V štandardnej fyzike sa hybnosť zachováva pre všetky jednotlivé častice a ich interakcie; podľa matematickej indukcie sa hybnosť zachováva pre fyzikálne systémy ľubovoľnej veľkosti. Ak si viete predstaviť, ako do seba dve častice narazia a vždy sa rozletia s rovnakou celkovou hybnosťou ako začínali, potom vidíte, že zmena škály z častíc na gigantickú komplikovanú zbierku ozubených kolies na tom nič nezmení. Aj keď zahrniete bilión biliárd atómov, 0 + 0 + … + 0 = 0.

Ale zákon zachovania energie, ako taký, nezabraňuje premene tepla na prácu. Môžete v skutočnosti postaviť uzavretú krabicu, ktoré premieňa ľadové kocky a uloženú elektrinu na teplú vodu. Ani to nie je ťažké. Energiu nemožno vytvoriť ani zničiť: Čistá zmena energie pri transformovaní (ľadové kocky + elektrina) na (teplá voda) musí byť 0. Takže by to neporušilo zákon zachovania energie, ako taký, keby ste to urobili naopak...

Stroje večného pohybu druhého typu, ktoré premieňajú teplú vodu na elektrický prúd a ľadové kocky, zakazuje *druhý* zákon termodynamiky.

Druhý zákon je trochu ťažší na porozumenie, a je vo svojej podstate bayesiánsky.

Áno, naozaj. Základný *fyzikálny* zákon za druhým zákonom termodynamiky je veta, ktorú možno dokázať v štandardnom modeli fyziky: *Pri vývoji ľubovoľného uzavretého systému v čase sa zachováva objem fázového priestoru.*

Predstavte si, že držíte vysoko nad zemou loptu. Tento stav vecí môžeme opísať ako bod v mnohorozmernom priestore, kde aspoň jeden z rozmerov je „výška lopty nad zemou". Potom, keď loptu pustíte, sa hýbe, a takisto aj ten bezrozmerný bod vo fázovom priestore, ktorý opisuje celý systém

→ http://lesswrong.com/lw/gv/outside_the_laboratory/

zahŕňajúci aj vás aj loptu. „Fázový priestor“ v jazyku fyziky znamená, že sú tam rozmery pre hybnosť častíc, nielen pre ich polohu, čiže systém 2 častíc by mal 12 rozmerov, 3 rozmery pre polohu každej častice a 3 rozmery pre hybnosť každej častice.

Keby ste mali mnohorozmerný priestor, ktorého každý rozmer by opisoval polohu ozubeného kolesa v obrovskom systéme ozubených kolies, potom keby ste roztočili tieto kolesá, tento bod v mnohorozmernom fázovom priestore by sa rozbehol a skákal by hore-dole. Čo znamená, že ak si môžete predstaviť veľký a zložitý stroj ako jediný bod v mnohorozmernom priestore, tak si viete aj predstaviť fyzikálne zákony opisujúce správanie tohto stroja v čase ako opisujúce trasu tohto bodu cez fázový priestor.

Druhý zákon termodynamiky je dôsledkom vety, ktorú možno dokázať v štandardnom modeli fyziky: Ak si vezmete nejaký objem fázového priestoru a vyvíjate ho dopredu v čase pomocou štandardnej fyziky, celkový objem fázového priestoru sa zachováva.

Napríklad:

Majme dva systémy, X a Y, kde X má 8 možných stavov a Y má 4 možné stavy a spojený systém (X, Y) má 32 možných stavov.

Vývoj spojeného systému v čase možno opísať ako pravidlo, ktoré počiatočným bodom priraďuje budúce body. Napríklad systém môže začať v X7Y2, potom sa vyvinúť (podľa nejakej množiny fyzikálnych zákonov) do stavu X3Y3 o minútu neskôr. Čo znamená: keby X začalo na 7, a Y začalo na 2, a sledovali by sme ich 1 minútu, uvideli by sme, že X ide do 3 a Y ide do 3. Také sú zákony fyziky.

Ďalej, vykrojme si z tohto spojeného systému podpriestor S. S bude podpriestor ohraničený tým, že X je v stave 1 a Y je v stavoch 1-4. Celkový objem S je teda 4 stavy.

A predpokladajme, že podľa zákonov fyziky, ktoré riadia (X, Y), sa počiatočné stavy S správajú takto:

X1Y1 -> X2Y1
X1Y2 -> X4Y1
X1Y3 -> X6Y1
X1Y4 -> X8Y1

Takto v skratke funguje chladnička.

Podsystém X začal v úzkom rozsahu priestoru stavov – vlastne v jedinom stave 1 – a Y začal rozdelený v širšom rozsahu stavov, v stavoch 1-4. Po vzájomnej interakcii Y prešiel do úzkeho rozsahu a X skončil v širokom rozsahu; *ale celkový objem fázového priestoru sa zachoval*. 4 počiatočné stavy sa zmenili na 4 koncové stavy.

Je jasné, že dokiaľ fyzika v čase zachováva celkový objem fázového priestoru, nemôžete stlačiť Y viac než sa X rozšíri, a naopak – pre každý podsystém, ktorý natlačíte do užšieho rozsahu priestoru stavov, musí sa nejaký iný podsystém rozšíriť do širšieho rozsahu priestoru stavov.

Predpokladajme teraz, že sme si *neistí* ohľadom spojeného systému (X, Y) a že našu *neistotu* opisuje rovnomerné rozdelenie v S. To znamená, sme si celkom istí, že X je v stave 1, ale Y má rovnakú šancu byť v ľubovoľnom zo stavov 1-4. Ak na minútu zavrieme oči a potom ich opäť otvoríme, budeme očakávať Y v stave 1, ale X môže byť v ľubovoľnom zo stavov 2-8. V skutočnosti môže byť X iba v *niektorých* zo stavov 2-8, ale bolo by príliš drahé myslieť na to, v ktorých presne z týchto stavov môže byť, preto iba povieme 2-8.

Ak si vezmete Shannonovu entropiu našej neistoty ohľadom X a Y ako jednotlivých systémov, X začalo s 0 bitmi entropie, pretože malo jediný jasne určený stav, a Y začalo s 2 bitmi entropie, pretože malo rovnakú šancu byť v ľubovoľnom zo 4 možných stavov. (Nebola žiadna vzájomná informácia medzi X a Y.) Stalo sa trochu fyziky a hľa, entropia Y klesla na 0, ale entropia X vzrástla na log2(7) = 2,8 bitu. Entropia sa teda presunula z jedného systému do druhého, a znížila sa *v rámci* podsystému Y; ale vďaka nákladnosti zaznamenávania sme nesledovali niektoré informácie a preto sa (z nášho pohľadu) celková entropia zvýšila.

Keby existoval fyzikálny proces, ktorý by minulým stavom priraďoval budúce stavy takto:

X2,Y1 -> X2,Y1

X2,Y2 -> X2,Y1
X2,Y3 -> X2,Y1
X2,Y4 -> X2,Y1

Potom by ste mohli mať fyzikálny proces, ktorý by naozaj *znižoval entropiu*, pretože bez ohľadu na to, kde by ste začali, by ste skončili na rovnakom mieste. Zákony fyziky vyvíjajúce sa v čase by stláčali fázový priestor.

Existuje však veta, Liouvilleova veta, ktorú možno dokázať pre naše zákony fyziky, ktorá hovorí, že toto sa nikdy nestane: fázový priestor sa zachováva.

Druhý zákon termodynamiky je dôsledkom Liouvilleovej vety: bez ohľadu na to, aká chytrá je vaša zostava kolies a ozubených kolies, nikdy nedokážete zmenšiť entropiu v jednom podsystéme bez jej zvýšenia niekde inde. Keď sa fázový priestor jedného podsystému zužuje, fázový priestor iného podsystému sa musí rozšíriť, a spoločný priestor si zachováva rovnaký objem.

Akurát to, čo bolo na začiatku *kompaktným* fázovým priestorom, si môže vyvinúť ohyby, slučky a vývrtky; takže aby ste nakreslili jednoduchú hranicu okolo celého tohto neporiadku, musíte nakresliť omnoho väčšiu hranicu než predtým – toto vytvára dojem, že sa entropia zvyšuje. (A v kvantových systémoch, kde rôzne vesmíry idú rôznym smerom, entropia naozaj rastie v každom lokálnom vesmíre. Ale túto komplikáciu zatiaľ vynechám.)

Druhý zákon termodynamiky je v skutočnosti vo svojej podstate pravdepodobnostný – ak sa opýtate na pravdepodobnosť, že horúca voda spontánne prejde do stavu „studená voda a elektrina“, taká pravdepodobnosť existuje, ale je veľmi malá. To neznamená, že Liouvilleova veta môže byť porušená s malou pravdepodobnosťou; predsa len, je to fyzikálna veta. Znamená to, že ak ste na začiatku vo veľkom objeme fázového priestoru, ale *neviete kde*, môžete priradiť maličkú *pravdepodobnosť* tomu, že skončíte v určitom konkrétnom objeme fázového priestoru. *Podľa vašich doterajších informácií*, s infinitezimálnou pravdepodobnosťou sa tento konkrétny pohár horúcej vody môže spontánne premeniť na elektrický prúd a ľadové kocky. (Zanedbávajúc, ako obyčajne, kvantové javy.)

Druhý zákon teda *je* od podstaty bayesiánsky. Keď príde na hocijaký skutočný termodynamický systém, je to prísne zákonité tvrdenie o vašich *vedomostiach o* tomto systéme, ale iba pravdepodobnostné tvrdenie o systéme samotnom.

„Zadrž,“ poviete. „To nie je to, čo som sa učil na hodinách fyziky. Na našich hodinách som počul, že termodynamika je, veď vieš, o *teplotách*. Neistota je subjektívny stav mysle! Teplota pohára vody je objektívnou vlastnosťou vody! Čo má teplo spoločné s pravdepodobnosťou?“

Ach, vy čo máte málo dôvery.

V jednom smere je súvislosť medzi teplom a pravdepodobnosťou pomerne priamočiara: Ak jediná vec, ktorú viete o pohári vody, je jeho teplota, potom ste si omnoho viac neistí o pohári horúcej vody než o pohári studenej vody.

Teplo je odrážanie sa mnohých maličkých molekúl; čím sú horúcejšie, tým rýchlejšie idú. Nie všetky molekuly horúcej vody cestujú rovnakou rýchlosťou - „teplota“ nie je rovnomerná rýchlosť všetkých molekúl, je to priemerná rýchlosť molekúl, ktorá zodpovedá predpovedateľnému štatistickému rozdeleniu rýchlostí molekúl – každopádne, pointa je, že čím horúcejšia je voda, tým rýchlejšie sa molekuly vody *môžu* pohybovať, a preto máte viac neistoty o rýchlosti (ako vektore) ľubovoľnej *jednotlivej* molekuly. Keď vynásobíte dokopy vaše neistoty o všetkých jednotlivých molekulách, budete mať *exponenciálne* viac neistoty o celom pohári vody.

Vezmete logaritmus tohto exponenciálneho množstva neistoty a nazveme to entropia. Takže to celé funguje, ako vidíte.

Spojenie v opačnom smere je menej zrejmé. Predstavte si, že je pohár vody, o ktorom ste na začiatku vedeli iba to, že jeho teplota je 72 stupňov. Náhle vám svätý Laplace zjaví presné polohy a rýchlosti všetkých atómov vody. Teraz dokonale poznáte stav vody, takže podľa informačno teoretickej definície entropie je jeho entropia nulová. Robí to aj jeho *termodynamickú* teplotu nulovou? Je voda studenšia preto, lebo o nej vieme viac?

Na chvíľku ignorujúc kvantovitosť, odpoveď je: Áno! Je to tak!

Maxwell sa raz opýtal: Prečo nemôžeme vziať rovnomerne horúci plyn, rozdeliť ho na dve oddelenia A a B, a nechať iba rýchlo letiace molekuly prejsť z B do A, a iba pomaly letiace molekuly z A do B? Keby ste vedeli postaviť takúto prepážku, o chvíľu by ste mali horúci plyn na strane A a studený plyn na strane B. To by bol lacný spôsob, ako chladiť potraviny, nie?

Činiteľ, ktorý skúma každú molekulu plynu a rozhoduje, či ju nechá prejsť, je známy ako „Maxwellov démon“. A dôvod, prečo nemôže takýmto spôsobom postaviť fungujúcu chladničku je, že Maxwellov démon generuje entropiu v procese skúmania molekúl plynu a rozhodovania sa, ktorú z nich nechá prejsť.

Ale predpokladajme, že by ste už *vedeli*, kde sú všetky molekuly plynu?

Potom by ste naozaj *mohli* spustiť Maxwellovho démona a vyťažiť užitočnú prácu.

Takže (opäť na chvíľu ignorujúc kvantové javy) keby ste *poznali* stavy všetkých molekúl v pohári horúcej vody, bola by studená v skutočnom termodynamickom zmysle: mohli by ste z nej vybrať všetku elektrinu a zanechať ľadovú kocku.

Toto neporušuje Liouvilleovu vetu, pretože ak Y je voda a *vy* ste Maxwellov démon (označený M), fyzikálny proces sa správa ako:

M1,Y1 -> M1,Y1

M2,Y2 -> M2,Y1

M3,Y3 -> M3,Y1

M4,Y4 -> M4,Y1

Pretože Maxwellov démon *pozná* presný stav Y, toto je vzájomná informácia medzi M a Y. Vzájomná informácia znižuje spoločnú entropiu (M, Y): H(M, Y) = H(M) + H(Y) – I(M, Y). M má 2 bity entropie, Y má 2 bity entropie, a ich vzájomná informácia sú 2 bity, preto (M, Y) má dokopy 2 + 2 – 2 = 2 bity entropie. Tento fyzikálny proces iba premieňa „chladnosť“ (negentropiu) vzájomnej informácie na skutočne chladnú vodu – na konci má M 2 bity entropie, Y má 0 bitov entropie, a vzájomná informácia je 0. Na tom nie je nič zlé!

A nehovorte mi, že vedomosti sú „subjektívne“. Vedomosti musia byť reprezentované v nejakom mozgu, a to ich robí rovnako fyzikálnymi ako hocičo iné. Aby M fyzicky reprezentovalo presný obraz stavu Y, musí fyzikálny stav M korelovať so stavom Y. Môžete z toho ťažiť termodynamické výhody – volá sa to Szilardov stroj.

Alebo ako to povedal E. T. Jaynes: „Staré porekadlo ‚poznanie je sila‘ je veľmi výstižná pravda, aj v ľudských vzťahoch, aj v termodynamike.“

A naopak, *jeden podsystém nemôže zvýšiť vzájomnú informáciu s iným podsystémom bez (a) interakcie s ním a (b) vykonávania termodynamickej práce.*

Inak by ste mohli postaviť Maxwellovho démona a porušiť druhý zákon termodynamiky – čím by ste následne porušili Liouvilleovu vetu – čo je zakázané štandardným modelom fyziky.

Čo znamená: **Aby ste si o niečom utvorili presné názory, *naozaj* to musíte pozorovať.** Je to celkom fyzikálny, celkom skutočný proces: každá rozumná myseľ koná „prácu“ v termodynamickom zmysle, nielen v zmysle myšlienkového úsilia.

(Občas sa hovorí, že táto termodynamická práca spočíva vo vymazávaní bitov pri ich príprave na ďalšie pozorovanie – ale toto rozlišovanie je iba otázkou slov a uhla pohľadu; matematika je jednoznačná.)

(Objavovanie logických „právd“ je komplikácia, ktorú zatiaľ nebudem brať do úvahy – prinajmenšom preto, lebo zatiaľ sám rozmýšľam o presnom formalizme. V termodynamike sa poznanie logických právd nepočíta ako negentropia; čo by sa dalo očakávať, pretože reverzibilný počítač môže počítať logické pravdy pri ľubovoľne nízkej cene. Všetko, čo som tu povedal, platí pre logicky vševediacu myseľ: ľubovoľná menšia myseľ bude nevyhnutne menej efektívna.)

„Utvárať si správne názory si vyžaduje zodpovedajúce množstvo indície“ je veľmi výstižná pravda aj v ľudských vzťahoch, aj v termodynamike: keby slepá viera naozaj fungovala ako metóda vyšetrovania, mohli by ste premieňať teplú vodu na elektrinu a ľadové kocky. Jednoducho postavte Maxwellovho démona, ktorý má slepú vieru v rýchlosti molekúl.

Stroje na poznanie nie sú také odlišné od tepelných strojov, hoci narábajú s entropiou jemnejším spôsobom než je pálenie benzínu. Napríklad do tej miery, nakoľko je stroj na poznanie neefektívny, musí vyžarovať odpadové teplo, tak ako motor auta alebo chladnička.

„Chladná racionalita“ je pravdivá v zmysle, o ktorom sa hollywoodskym scenáristom ani nesnívalo (a nepravdivá v zmysle, o ktorom sa im snívalo).

Takže pokiaľ mi nepoviete, ktorý *konkrétny krok* vášho argumentu porušuje zákony fyziky tým, že vám dáva pravdivú vedomosť o nevidenom, nečakajte, že uverím, že to dokáže nejaký veľký komplikovaný chytrý argument.

*→

187. Názory ako večný pohyb

Včerajší článok končil:

> **Aby ste si o niečom utvorili presné názory, *naozaj* to musíte pozorovať.** Je to celkom fyzikálny, celkom skutočný proces: každá rozumná myseľ koná „prácu“ v termodynamickom zmysle, nielen v zmysle myšlienkového úsilia. ... Takže pokiaľ mi nepoviete, ktorý *konkrétny krok* vášho argumentu porušuje zákony fyziky tým, že vám dáva pravdivú vedomosť o nevidenom, nečakajte, že uverím, že to dokáže nejaký veľký komplikovaný chytrý argument.

Jedno z hlavných ponaučení z matematickej analógie medzi termodynamikou a poznaním je, že z obmedzení pravdepodobnosti sa nedá uniknúť; pravdepodobnosť môže byť „subjektívnym stavom názoru“, ale zákony pravdepodobnosti sú pevnejšie než oceľ.

Ľudia sa v tradičnom školstve naučia, že učiteľ vám povie nejaké veci, a vy im musíte veriť a odrecitovať ich naspäť; ale ak nejaký názor navrhne iba *študent*, nemusíte to poslúchnuť. Mapujú si oblasť názorov na oblasť autority, a myslia si, že istý názor je ako príkaz, ktorý treba poslúchnuť, ale pravdepodobnostný názor je iba odporúčanie.

Potom sa pozrú na žreb lotérie a povedia: „Ale nemôžeš mi *dokázať*, že nevyhrám, však?“ Čím myslia: „Možno si vypočítal nízku pravdepodobnosť výhry, ale keďže je to *pravdepodobnosť*, je to iba *odporúčanie*, a ja *môžem* veriť, čomu sa mi zachce.“

Tu je malý pokus: Hoďte vajce o zem. Pravidlo, ktoré hovorí, že toto vajce sa spontánne neposkladá a nevyskočí vám do ruky, je iba pravdepodobnostné. Iba odporúčanie, ak chcete. Zákony termodynamiky sú pravdepodobnostné, takže to nemôžu byť naozaj *zákony* v rovnakom zmysle ako je zákon „nezabiješ“... hej?

Prečo teda toto odporúčanie skrátka neignorovať. Potom sa to vajce samo poskladá... hej?

Možno vám pomôže myslieť na to takto – ak máte stále nejakú vytrvalú intuíciu, že neisté názory nie sú autoritatívne:

V skutočnosti môže existovať veľmi malá šanca, že sa toto vajce spontánne poskladá. Ale nemôžete *očakávať*, že sa poskladá. *Musíte* očakávať, že sa rozbije. Váš povinný názor je, že pravdepodobnosť spontánneho poskladania vajca je ~0. Pravdepodobnosti nie sú istoty, ale *zákony* pravdepodobnosti sú matematické vety.

Ak o tom pochybujete, skúste hodiť vajce na zem niekoľko deciliónkrát, ignorujte odporúčania termodynamiky a očakávajte, že sa spontánne poskladá, a uvidíte, čo sa stane. Pravdepodobnosti môžu byť subjektívne stavy názoru, ale zákony, ktorými sa riadia, sú pevnejšie než oceľ. Raz som poznal človeka, ktorý bol *presvedčený*, že jeho systém kolies a ozubených kolies vytvorí nereaktívny pohon, a mal excelovský zošit, ktorý to dokazoval – ktorým nám samozrejme nemohol ukázať, lebo zatiaľ ešte stále tento systém vyvíjal. V klasickej mechanike je porušenie zákona zachovania hybnosti *dokázateľne* nemožné. Takže ľubovoľný excelovský zošit vypočítaný *podľa pravidiel klasickej mechaniky* musí

→ http://lesswrong.com/lw/o5/the_second_law_of_thermodynamics_and_engines_of/

nevyhnutne ukazovať, že nereaktívny pohon neexistuje – pokiaľ váš stroj nie je natoľko zložitý, že ste urobili chybu vo výpočte.

A podobne, keď polovycvičení alebo z desatiny vycvičení racionalisti opustia svoje umenie a skúsia uveriť bez dôkazov iba tento jedinýkrát, často postavia rozľahlé budovy zdôvodnenia, zmätú sami seba akurát natoľko, aby zakryli magické kroky.

Môže byť veľmi otravné hľadať, kde presne nastáva táto mágia – ich štruktúra argumentu má sklon meniť sa a uhýbať, ako ich vypočúvate. Ale vždy je tam nejaký krok, kde sa maličká pravdepodobnosť premení na veľkú – kde sa pokúšajú veriť bez dôkazu – kde vstúpia do neznáma s myšlienkou: „Nikto mi nemôže dokázať, že sa mýlim."

Ich noha prirodzene stúpi do vzduchu, pretože v krajine Možnosti je omnoho viac vzduchu než zeme. Ach, ale v Možnosti *existuje* (exponenciálne malé) možstvo zeme, a vy máte (exponenciálne malú) pravdepodobnosť, že na ňu stúpite náhodou, takže možno tentokrát vaša noha *stúpi* na správne miesto! Je to *iba* pravdepodobnosť, preto to musí byť *iba* odporúčanie.

Presný stav pohára vriacej vody nemusíte poznať – vaša nevedomosť jeho presného stavu je dokonca to, čo robí z kinetickej energie molekúl „teplo" namiesto práce, ktorá čaká na vyťaženie ako hybnosť otáčajúceho sa zotrvačníka. Takže voda by *mohla* ochladiť vašu ruku namiesto jej zohriatia, s pravdepodobnosťou ~0.

Rozhodnite sa ignorovať zákony termodynamiky a strčte tam napriek tomu ruku, a popálite sa.

„Ale to nemôžeš vedieť!"

Neviem to naisto, ale je *povinné*, aby som *očakával*, že sa to stane. Pravdepodobnosti nie sú logické pravdy, ale *zákony* pravdepodobnosti áno.

„Ale čo ak budem hádať stav vriacej vody a čo ak uhádnem správne?"

Vaša šanca uhádnuť správne šťastnou náhodou je ešte *menšia* než šanca, že vriaca voda šťastnou náhodou vašu ruku ochladí.

„Ale nemôžeš mi *dokázať*, že neuhádnem správne."

Môžem (dokonca musím) tomu priradiť extrémne nízku pravdepodobnosť.

„Lenže to nie je to isté ako istota."

Hej, možno keď do svojho argumentu pridáte dostatok kolies a ozubených kolies, bude premieňať horúcu vodu na elektrinu a kocky ľadu! Alebo aspoň nebudete vidieť dôvod, prečo by to tak *nemohlo* byť.

„Správne! Ja *nevidím* dôvod, prečo by to tak nemohlo byť! Takže možno je to tak!"

Ďalšie ozubené koleso? To len robí váš stroj ešte *menej* efektívnym. Ak z neho nebolo perpetuum mobile už predtým, každé ozubené koleso navyše to robí ešte menej účinným.

Každá dodatočná podrobnosť vo vašom argumente nevyhnutne znižuje jeho celkovú pravdepodobnosť. Pravdepodobnosť, že ste porušili druhý zákon termodynamiky, a ani neviete presne ako, uhádnutím presného stavu vriacej vody bez indície, takže do nej môžete strčiť prst a nepopálite sa, je nevyhnutne ešte nižšia než pravdepodobnosť, že strčíte prst do vriacej vody a nepopálite sa.

Toto všetko hovorím, lebo ľudia naozaj stavajú tieto rozľahlé budovy argumentov v rámci viery bez dôkazov. Človek sa musí naučiť vidieť analógiu medzi tým a všetkými kolesami a ozubenými kolesami, ktoré dotyčný pridáva do svojho nereaktívneho pohonu, dokiaľ konečne nenazbiera dostatok komplikácií na to, aby sa v jeho excelovom zošite vyskytla chyba.

*→

188. Hľadanie bayesovskej štruktúry

> „*Gnómske prilby* by nemali fungovať. Ich samotná konštrukcia napohľad popiera podstatu thaumaturgického zákona. V skutočnosti sú nemožné. Ako väčšina produktov gnómskych myslí, zahŕňajú veľké množstvo zvončekov a píšťaliek, a veľmi málo podstaty.

→ http://lesswrong.com/lw/o6/perpetual_motion_beliefs/

> Tie, ktoré fungujú, majú zvyčajne v sebe *menšiu prilbu*, ktorá je vždy skrytá a zamaskovaná, aby vyzerala neškodne a nepodstatne.“
>
> --prostredie kampane Spelljammer

Videli sme, že poznanie znamená vzájomnú informáciu medzi mysľou a jej prostredím, a videli sme, že táto vzájomná informácia je negentropia v skutočnom fyzikálnom zmysle: Ak viete, kde sú molekuly a ako rýchlo sa hýbu, môžete premeniť teplo na prácu pomocou Maxwellovho démona / Szilardovho stroja.

Videli sme, že vytvorenie si pravdivých názorov bez indície je rovnaký druh nepravdepodobnosti ako že sa horúca voda spontánne usporiada na kocky ľadu a elektrinu. Rozumnosť si vyžaduje „prácu“ v termodynamickom zmysle, nie iba v zmysle myšlienkového úsilia; myseľ musí vyžarovať teplo, ak nie je dokonale efektívna. Táto poznávacia práca sa riadi teóriou pravdepodobnosti, ktorej špeciálnym prípadom je termodynamika. (Štatistická mechanika je špeciálny prípad štatistiky.)

Keby ste videli stroj, ktorý stále otáča kolesom bez zrejmého pripojenia do zásuvky alebo iného zdroja energie, potom by ste hľadali skrytú batériu alebo blízky vysielač energie – niečo, čo vysvetlí vykonávanú prácu bez porušenia zákonov fyziky.

Ak teda myseľ dochádza k pravdivým názorom a predpokladáme, že nebol porušený druhý zákon termodynamiky, táto myseľ musí robiť niečo aspoň *približne* bayesovské – prinajmenšom jeden proces s istým druhom bayesovskej štruktúry *niekde* – inak by to *nemohlo fungovať*.

Na začiatku, v čase T = 0, nemá myseľ žiadnu vzájomnú informáciu s podsystémom S vo svojom prostredí. V čase T = 1 má myseľ 10 bitov vzájomnej informácie s S. Niekde medzi tým sa myseľ musela stretnúť s indíciou – podľa bayesovskej definície indície, pretože každá bayesovská indícia je vzájomná informácia, a každá vzájomná informácia je bayesovská indícia, sú to len rôzne spôsoby, ako sa na to pozerať – a spracovala aspoň trochu z tej indície, akokoľvek neefektívne, v správnom smere podľa bayesa, prinajmenšom občas. Táto myseľ sa musela *pohnúť v harmónii s Bayesom* aspoň trochu, niekde v tom procese – buď to, alebo porušila druhý zákon termodynamiky vytvorením vzájomnej informácie z ničoho.

V skutočnosti, ľubovoľná *časť* kognitívneho procesu, ktorý *užitočne prispieva* k hľadaniu pravdy, musí mať aspoň trochu bayesovskej štruktúry – musí byť v harmónii s Bayesom, v nejakom bode – musí čiastočne zodpovedať bayesovskému toku, akokoľvek šumovo – napriek hocijakému množstvu maskujúcich zvončekov a píšťaliek – dokonca aj keď je táto bayesovskú štruktúra zrejmá iba v kontexte okolitých procesov. Inak by nemohla *pomáhať*.

Ako sa filozofi zamýšľali nad podstatou slov! Koľko atramentu sa minulo na správnu definíciu slov, a správny význam definície, a správny význam významu! Koľko zbierok kolies a ozubených kolies postavili vo svojich vysvetleniach! A po celý čas to bola skrytá forma bayesovského odvodzovania!

Naozaj som bol trochu sklamaný, že nikto v obecenstve nevyskočil a nepovedal: „*Áno! Áno, to je ono! Samozrejme! Po celý čas to bol Bayes!*“

Ale možno nie je *až* také vzrušujúce vidieť niečo, čo na povrchu *nevyzerá* bayesovsky, odhalené ako Bayes v chytrom maskovaní, ak: (a) neodhalíte toto tajomstvo sami, ale čítate o tom, ako to urobil niekto iný (Newtona bavilo počítanie integrálov viac než väčšinu študentov), a (b) neuvedomíte si, že *hľadanie skrytej bayesovskej štruktúry* je táto obrovská, náročná, všadeprítomná úloha, ako hľadanie Svätého Grálu.

Je to rôzna výprava pre každý aspekt poznávania, ale Grál sa nakoniec vždy *ukáže* byť ten istý. Musí to však byť ten *správny* Grál – a *celý* Grál, bez chýbajúcich častí – takže musíte vždy ísť na výpravu za celou odpoveďou, nech má *akúkoľvek* podobu, namiesto snahy umelo vybudovať približne plusmínus grálovité argumenty. *Potom* nakoniec vždy nájdete ten istý Svätý Grál.

Už mi bolo vytýkané, že dlhými článkami možno strácam niektorých zo svojich čitateľov, pretože „nedávam jasne najavo, kam smerujem“...

...lenže nie je také jednoduché ľuďom skrátka povedať, kam idete, keď idete niekam, ako je *toto*.

Nie je veľmi užitočné iba *vedieť, že* nejaká forma poznania je bayesovká, ak *neviete, ako* je bayesovská. Ak nedokážete uvidieť podrobný tok pravdepodobnosti, nemáte nič, iba heslo – alebo, trochu zhovievavejšie, náznak, aký tvar by mohla mať odpoveď; ale rozhodne nemáte odpoveď. Preto existuje veľká výprava za skrytou bayesovskou štruktúrou, namiesto iba povedania „Bayes!“ a hotovo. Bayesovská štruktúra sa môže schovávať rôzne preoblečená, maskovaná, skrytá za húštinou kolies a ozubených kolies, zakrytá zvončekmi a píšťalkami.

Výpravu za Svätým Bayesom začnete chápať tak, že sa dozviete o poznávacom jave XYZ, ktorý vyzerá naozaj užitočne – a je tu partia filozofov, ktorí o jeho pravej podstate debatujú celé stáročia, a stále neskončili – a je tu partia vedcov UI, ktorí sa snažia dosiahnuť, aby to robil počítač, ale tiež sa nevedia dohodnúť na filozofii...

A – *ha, to je divné!* - tento poznávací jav sa na povrchu nepodobá na nič bayesovské, ale pod tým je nejasná štruktúra, ktorá má bayesovské vysvetlenie – ale počkajte, stále sa tam deje niečo užitočné, čo sa nedá vysvetliť bayesovskými pojmami – nie, počkajte, aj to *je* bayesovské – ACH BOŽE, tento *celkom odlišný* poznávací proces, ktorý *tiež* na povrchu nevyzeral bayesovsky, MÁ TIEŽ BAYESOVSKÚ ŠTRUKTÚRU – vydržte, *robia* tieto nebayesovské časti vôbec niečo?

- Áno: Ach, tie sú tiež bayesovské!
- Nie: Nebesá, aký hlúpy dizajn. Mohol by som zjesť za vedro aminokyselín a vyzvracať lepšiu mozgovú architektúru než je toto.

Keď sa vám toto stane niekoľkokrát, začnete akosi chytať rytmus. O tom tu hovorím, o rytme.

Pokúšať sa hovoriť o rytme je trochu ako pokúšať sa tancovať o architektúre.

To ma necháva trochu v kaši, keď ide o vysvetľovanie vopred, kam smerujem. Viem zo skúsenosti, že ak poviem: „Tajomstvom vesmíru je Bayes,“ niektorí ľudia možno povedia: „Áno! Tajomstvom vesmíru je Bayes!“; a iní vyprsknú a povedia: „Aký si obmedzený; pozri na všetky tieto zvyšné ad-hoc ale úžasne užitočné metódy, ako napríklad regularizovaná lineárna regresia, ktoré mám vo svojej zbierke nástrojov.“

Dúfal som, že ak budem mať poruke konkrétny príklad „niečoho, čo na povrchu vôbec nevyzeralo bayesovsky, ale nakoniec sa ukázalo byť bayesovské“ - *a* vysvetlenie rozdielu medzi heslom a vedomosťou – *a* vysvetlenie rozdielu medzi nástrojmi a zákonmi – *potom* možno dokážem úspešne sprostredkovať toľko z toho rytmu, koľko sa len dá pochopiť bez osobnej účasti na výprave.

Toto samozrejme nie je *celé* tajomstvo bayesovskej konšpirácie, ale je to všetko, čo dokážem v túto chvíľu odovzdať. Navyše, celé tejomstvo pozná iba bayesovská rada, a keby som vám ho povedal, musel by som vás zamestnať.

Aby ste *prekukli* povrchnú nesúvislosť poznávacieho procesu až na bayesovskú štruktúru *pod ním* – aby ste vnímali toky pravdepodobnosti, aby ste *vedeli ako*, nielen *vedeli že* aj toto poznanie je bayesovské – ako vždy je – ako vždy musí byť – aby ste dokázali cítiť Silu, na ktorej je založené všetko poznanie – toto je videnie Bayesa.

> „...A kráľovná Kašfy vidí pomocou Hadieho Oka.“
>
> „Neviem, či pomocou neho vidí,“ povedal som. „Stále sa zotavuje z operácie. Ale je to zaujímavá myšlienka. Keby ním mohla vidieť, na čo by mohla pozerať?“
>
> „Číre, chladé čiary večnosti, trúfam si povedať. Pod všetkým Tieňom.“
>
> --Roger Zelazny, Princ chaosu[191]

*→

191 Roger Zelazny, *Prince of Chaos* (Thorndike Press, 2001).

→ http://lesswrong.com/lw/o7/searching_for_bayesstructure/

P: Redukcionizmus pre začiatočníkov

189. Rozpustenie otázky

„Ak strom padne v lese, ale nikto ho nepočuje, vydá pritom zvuk?“

Neodpovedal som na túto otázku. Nevybral som si postoj „Áno!“ alebo „Nie!“ a nezačal som ho obhajovať. Namiesto toho som odišiel a rozobral som ľudský algoritmus na spracovávanie slov, zašiel som dokonca tak ďaleko, že som načrtol obrázok neurónovej siete. Na konci, dúfam, nezostala žiadna otázka – dokonca ani pocit otázky.

Mnohí filozofi – najmä amatérski filozofi a starovekí filozofi – majú spoločný nebezpečný inštinkt: ak im dáte otázku, pokúsia sa na ňu odpovedať.

Povedzte napríklad: „Máme slobodnú vôľu?“

Nebezpečným inštinktom filozofie je začať zoraďovať argumenty v prospech a argumenty v neprospech, zvážiť ich, uverejniť ich v prestížnom filozofickom časopise, a nakoniec uzavrieť: „Áno, musíme mať slobodnú vôľu“ alebo „Nie, nemôžeme mať slobodnú vôľu“.

Niektorí filozofi sú dostatočne múdri, aby si pripomenuli varovanie, že väčšina filozofických debát sú v skutočnosti hádky o význame slova, alebo zmätok vytvorený používanám rôznych významov pre to isté slovo na rôznych miestach. Takže sa snažia veľmi presne definovať, čo myslia slovami „slobodná vôľa“ a potom sa opäť spýtajú: „Máme slobodnú vôľu? Áno alebo nie?“

A ešte múdrejší filozof môže mať podozrenie, že zmätok ohľadom „slobodnej vôle“ ukazuje, že samotný pojem je pochybený. Sledujú teda dráhu Tradičnej Rozumnosti: Argumentujú, že „slobodná vôľa“ je vnútorne rozporný pojem, alebo je nezmyselný, lebo nemá testovateľné dôsledky. A potom uverejnia tieto zdrvujúce pozorovanie v prestížnom filozofickom časopise.

Lenže *dokázať, že* ste zmätení, nemusí spôsobiť, že sa budete cítiť *menej* zmätene. Dokázať, že otázka je nezmyselná, nám nemusí pomôcť odpovedať na ňu.

Filozofov inštinkt je nájsť najlepšie brániteľnú pozíciu, uverejniť ju, a ísť ďalej. Lenže „naivný“ pohľad, inštinktívny pohľad, je faktom o ľudskej psychológii. Môžete dokazovať, že slobodná vôľa je nemožná, dokiaľ Slnko nevyhasne, ale zanecháva to nevysvetlený fakt kognitívnej vedy: Ak slobodná vôľa neexistuje, čo sa odohráva v hlave človeka, ktorý si myslí, že existuje? Toto nie je rečnícka otázka!

Je faktom o ľudskej psychológii, že ľudia si myslia, že majú slobodnú vôľu. Nájdenie lepšie obhániteľnej *filozofickej pozície* nemení ani nevysvetľuje tento *psychologický fakt*. Filozofia vás môže viesť k *odmietnutiu* tohto pojmu, ale odmietnutie pojmu nie je to isté ako pochopenie kognitívnych algoritmov za ním.

Môžete sa pozrieť na štandardnú debatu ohľadom „ak strom padne v lese a nikto to nepočuje, vzniká pritom zvuk?“ a urobiť vec Tradičnej Rozumnosti: pozorovať, že títo dvaja si neodporujú v žiadnom bode očakávanej skúsenosti, a triumfálne vyhlásiť, že hádka je zbytočná. To je v tomto konkrétnom prípade pravda; ale *ako kognitívni vedci* sa pýtame, prečo tí hádajúci sa vôbec takúto chybu urobili?

Kľúčová myšlienka programu heuristík a skreslení je, že *chyby*, ktoré robíme, často o našich základných poznávacích algoritmoch odhaľujú viac než naše správne odpovede. Preto (kladiem si raz za čas otázku), aký druh dizajnu mysle zodpovedá chybe hádania sa o padajúcich stromoch v opustených lesoch?

Kognitívne algoritmy, ktoré používame, *sú* tým, ako nám pripadá svet. A tieto kognitívne algoritmy nemusia zodpovedať skutočnosti v pomere jedna k jednej – ani len makroskopickej skutočnosti, niečo ešte skutočným kvarkom. V mysli môžu byť veci, ktoré skresľujú svet.

Napríklad v strede neurónovej siete môže byť visiaca jednotka, ktorá nezodpovedá žiadnej skutočnej veci, ani skutočnej vlastnosti skutočnej veci existujúcej niekde v skutočnom svete. Táto visiaca jednotka je často užitočná ako skratka pri výpočte, a preto ju máme. (Metaforicky povedané. Ľudská neurobiológia je iste omnoho zložitejšia.)

Táto visiaca jednotka nám *pripadá ako* nezodpovedaná otázka, dokonca aj po tom, čo sme odpovedali na všetky odpovedateľné otázky. Bez ohľadu na to, ako vám kto dokáže, že na tejto otázke nezávisí žiadna očakávaná skúsenosť, zostanete sa čudovať: „Ale vydá ten padajúci strom *naozaj* zvuk, alebo nie?“

Ale akonáhle si *podrobne* uvedomíte, ako váš mozog vytvára pocit otázky – akonáhle si uvedomíte, že váš pocit nezodpovedanej otázky zodpovedá iluzórnej centrálnej jednotke, ktorá chce vedieť, či má vyslať signál, ešte aj po tom, čo sme všetky krajné jednotky pripojili k známym hodnotám – alebo ešte lepšie, ak rozumiete, ako technicky funguje naivný Bayes – *potom* ste hotoví. Už nezostáva žiaden doznievajúci pocit zmätku, žiaden nejasný pocit nespokojnosti.

Ak doznieva *nejaký* pocit, že zostala nezodpovedaná otázka, alebo že ste boli do niečoho slovne prevalcovaní, je to znamenie, že ste túto otázku nerozpustili. Hmlistá nespokojnosť by mala byť rovnakým varovaním ako výkrik. *Naozaj* rozpustená otázka za sebou nezanecháva nič.

Triumfálne hromové vyvrátenie slobodnej vôle, absolútne nevyvrátiteľný dôkaz, že slobodná vôľa nemôže existovať, nám prípadá veľmi *uspokojujúco* – veľký potlesk pre domáci tím. A tak si môžete nevšimnúť, že – z pohľadu kognitívnej vedy – nemáte plné a uspokojivé popisné vysvetlenie toho, ako každý jednotlivý pocit vzniká, krok za krokom.

Nemusíte si dokonca ani chcieť pripustiť svoju nevedomosť z tohto pohľadu kognitívnej vedy, pretože by vám to pripadalo ako bod proti vášmu tímu. Uprostred rozbíjania všetkých hlúpych názorov o slobodnej vôli by mohlo znieť ako ústupok opačnej strane priznať, že ste niečo nechali nevysvetlené.

A tak možno prídete s narýchlo vymysleným evolučno psychologickým argumentom, že lovci a zberači, ktorí verili na slobodnú vôľu, mali s väčšou pravdepodobnosťou pozitívny pohľad na život, a tak sa rozmnožovali viac ako iní lovci a zberači – aby som dal len jeden príklad celkom falošného vysvetlenia. Ak poviete toto, *argumentujete, že* mozog vytvára ilúziu slobodnej vôle – ale *nevysvetľujete, ako*. Pokúšate sa zamietnuť opozíciu rozoberaním jej motívov – ale v príbehu, ktorý hovoríte, je ilúzia slobodnej vôle surový fakt. Nerozobrali ste túto ilúziu na kúsky, aby ste videli jej kolesá a ozubené kolesá.

Predstavte si, že pri štandardnej debate o strome padajúcom v opustenom lese najprv dokážete, že neexistujú rozdiely v očakávaní, a potom prídete s hypotézou: „Ale ľudia, ktorí hovorili, že debaty sú nezmyselné, boli možno vnímaní ako porazení, čím strácali spoločenské postavenie, takže teraz máme inštinkt hádať sa o význame slov.“ To je *argumentovanie že*, prípadne *vysvetľovanie prečo* nejaký zmätok existuje. Teraz sa pozrite na štruktúru neurónovej siete v kapitole Precíťte význam. To je *vysvetlenie ako*, rozobranie zmätku na menšie kúsky, ktoré samotné nie sú mätúce. Vidíte ten rozdiel?

Prísť s dobrými hypotézami o poznávacích algoritmoch (alebo aspoň s hypotézami, ktoré vydržia aspoň pol sekundy) je omnoho ťažšie než iba odmietnuť filozofický záver. Veru, je to celkom iné umenie. Pamätajte na toto, a mali by ste sa menej hanbiť povedať: „Viem, že čo hovoríš, nemôže byť pravda, a viem to dokázať. Neviem však nakresliť diagram ukazujúci, ako tvoj mozog robí túto chybu, takže ešte nie som hotový a budem pokračovať v skúmaní.“

Toto všetko hovorím, pretože sa mi občas zdá, že aspoň 20 % efektívnosti šikovného racionalistu v skutočnom svete vychádza z toho, že sa nezastaví príliš skoro. Ak sa budete ďalej pýtať, časom sa do svojho cieľa dostanete. Ak sa príliš zavčasu rozhodnete, že ste našli odpoveď, tak nie.

Najväčšou výzvou je všimnúť si, kedy ste zmätení – dokonca aj keď to vyzerá iba ako maličký kúsoček zmätku – a dokonca aj keď oproti vám stojí niekto, kto *trvá* na tom, že ľudia majú slobodnú vôľu, a *vyškiera* sa na vás, a fakt, že neviete *presne* ako fungujú poznávacie algoritmy, nemá *nič spoločné* s prenikavou pochabosťou jeho postoja...

Ale keď dokážete vyložiť poznávací algoritmus v dostatočnom detaile, aby ste mohli prejsť myšlienkovým procesom krok za krokom a opísať, ako vzniká každý intuitívny vnem – rozložiť zmätok na menšie kúsky, ktoré samotné nie sú mätúce – *potom* ste hotoví.

Vyvarujte sa teda *názoru*, že ste hotoví, keď jediné, čo máte, je púhe triumfálne vyvrátenie chyby.

Ale keď ste *naozaj* hotoví, *viete*, že ste hotoví. Rozpustiť otázku je nezameniteľný pocit – ak ho raz zažijete, a po jeho zážitku sa rozhodnete nenechať sa viac oklamať. Tí, ktorí snívajú, nevedia, že snívajú, ale keď sa zobudíte, viete, že ste sa zobudili.

Čím chcem povedať: Keď ste hotoví, viete, že ste hotoví, ale žiaľ, naopak to neplatí.

Tu je teda váš problém na domácu úlohu: Aký druh poznávacieho algoritmu, ako to vyzerá zvnútra, by vytvoril pozorovanú debatu o „slobodnej vôli“?

Vaša úloha nie je argumentovať, či ľudia majú slobodnú vôľu alebo nie.

Vaša úloha nie je argumentovať, či je slobodná vôľa zlučiteľná s determinizmom alebo nie.

Vaša úloha nie je argumentovať, že táto otázka je nesprávne formulovaná, alebo že celý ten pojem je vnútorne rozporný, alebo že nemá testovateľné dôsledky.

Nechcem po vás, aby ste si vymysleli evolučné vysvetlenie, ako sa ľudia, ktorí verili v slobodnú vôľu, rozmnožovali; ani aby ste rozprávali, že pojem slobodnej vôle sa podozrivo podobá na skreslenie X. To sú iba pokusy *vysvetliť, prečo* ľudia veria na „slobodnú vôľu“, nie *vysvetliť ako*.

Vaša domáca úloha je napísať výpis zásobníka vnútorného algoritmu ľudskej mysle, ako vytvára intuíciu, ktorá poháňa celú tú prekliatu filozofickú hádku.

Toto je jedna z prvých skutočných výziev, o ktoré som sa raz pokúsil ako ašpirujúci racionalista. Jedna z tých relatívne ľahších hádaniek. Snáď vám podobne pomôže.

*→

—

190. Nesprávne otázky

Keď myseľ krája naprieč vláknami skutočnosti, vytvára *nesprávne otázky* – otázky, ktoré sa nedajú zodpovedať *pomocou ich vlastných slov*, iba rozpustiť pochopením poznávacieho algoritmu, ktorý vytvoril *dojem* otázky.

Dobrým signálom, že máte do činenia s „nesprávnou otázkou“ je, keď si nedokážete ani len *predstaviť* nejaký konkrétny stav fungovania sveta, ktorý by na túto otázku odpovedal. Keď sa ani len nezdá *možné* na danú otázku odpovedať.

Vezmite si napríklad štandardnú debatu o definíciách, o strome padajúcom v opustenom lese. Je nejaký možný stav sveta – ľubovoľný stav vecí – ktorý by zodpovedal tomu, keď slovo „zvuk“ *naozaj znamená* akustické vibrácie, alebo keď *naozaj znamená* sluchové vnemy?

(„Ale áno,“ povie niekto, „je to stav vecí, kde ‚zvuk‘ znamená akustické vibrácie.“ Tak si zahrajme tabu so slovom „znamená“ a „reprezentuje“ a všetkými podobnými synonymami a opäť mi opíšte: Ako má svet byť, v akom stave vecí, aby jedna strana mala pravdu a druhá sa mýlila?)

Alebo ak vám toto pripadá ľahké, vezmite si slobodnú vôľu: Aký konkrétny stav vecí, či už v deterministickej fyzike alebo vo fyzike, ktorá obsahuje náhodný prvok, ktorý si hádže kockou, by mohol zodpovedať tomu, že máme slobodnú vôľu?

A ak sa vám aj *toto* zdá príliš ľahké, potom sa opýtajte: „Prečo vôbec niečo existuje?“ a potom mi povedzte, ako by vôbec mohla *vyzerať* uspokojujúca odpoveď na túto otázku.

A nie, nepoznám odpoveď na to posledné. *Viem* však uhádnuť jednu vec na základe mojich predchádzajúcich skúseností s nezodpovedateľnými otázkami. Tá odpoveď sa nebude skladať z nejakej veľkej triumfálnej Prvotnej Príčiny. Táto otázka odíde preč v dôsledku nejakého vhľadu, ako moje myšlienkové algoritmy skresľujú skutočnosť, po ktorom pochopím, ako bola samotná táto otázka od začiatku nesprávna – ako samotná otázka predpokladala nejaký klam, obsahovala skreslenie.

Tajomstvo existuje v mysli, nie v skutočnosti. Ak neviem o nejakom jave, je to fakt o mojom stave mysle, nie fakt o samotnom jave. Tobôž ak sa zdá, že odpoveď ani nemôže existovať: Zmätok existuje na mape, nie na území. *Nezodpovedateľné* otázky neoznačujú miesta, kde do vesmíru vstupuje mágia. Označujú miesta, kde vaša myseľ skresľuje skutočnosť.

→ http://lesswrong.com/lw/of/dissolving_the_question/

Takéto otázky *musíme* rozpustiť. Keď sa na ne pokúšate odpovedať, stávajú sa zlé veci. Nevyhnutne to vytvára ten najhorší druh Tajomnej Odpovede na Tajomnú Otázku: taký, kde prídete s napohľad silnými argumentmi pre svoju Tajomnú Odpoveď, ale táto „odpoveď“ vám neumožní robiť žiadne nové predpovede, dokonca ani v spätnom pohľade, a daný jav má stále tú istú posvätnú nevysvetliteľnosť akú mal na začiatku.

Mohol by som si napríklad tipnúť, že odpoveď na záhadu Prvej Príčiny je, že nič *neexistuje* – že samotný pojem „existencie“ je pomýlený. Ale keby ste tomu úprimne verili, boli by ste o niečo menej zmätení? Ja tiež nie.

Ale tá krásna vec na *nezodpovedateľných* otázkach je, že sú *vždy* riešiteľné, prinajmenšom podľa mojej skúsenosti. Čo si pomyslela kráľovná Alžbeta I. ako prvú vec ráno, keď sa zobudila na svoje štyridsiate narodeniny? Keďže si viem ľahko *predstaviť* odpovede na túto otázku, ľahko vidím, že na to asi nikdy nebudem vedieť *naozaj* odpovedať, pretože skutočná informácia sa stratila v čase.

Na druhej strane: „Prečo vôbec niečo existuje?“ znie tak dokonale nemožne, že dokážem vydedukovať, že som jednoducho zmätený, či už takým alebo onakým spôsobom, a že pravda pravdepodobne nie je až taká zložitá v absolútnom mysle, a keď raz zmätok odíde, dokážem ju uvidieť.

Toto môže vyzerať antiintuitívne, ak ste nikdy neriešili nezodpovedateľné otázky, ale ubezpečujem vás, že *takto* tie veci fungujú.

Nabudúce: Jednoduchý trik na zvládanie „nesprávnych otázok“.

*→

191. Napravenie nesprávnej otázky

Keď sa stretnete s *nezodpovedateľnou* otázkou – otázkou, na ktorú sa zdá nemožné čo len *predstaviť* si odpoveď – existuje jednoduchý trik, ktorý môže túto otázku premeniť na riešiteľnú.

Porovnajte:

- „Prečo mám slobodnú vôľu?“
- „Prečo si myslím, že mám slobodnú vôľu?“

Pekná vec na tej druhej otázke je, že má *zaručene* skutočnú odpoveď, či už niečo také ako slobodná vôľa *existuje alebo neexistuje*. Otázka: „Prečo mám slobodnú vôľu?“ alebo „Mám slobodnú vôľu?“ vás posiela myslieť na drobné detaily fyzikálnych zákonov, také vzdialené od makroskopickej úrovne, že ich nedokážete začať vnímať vlastnými očami. Navyše sa pýtate: „Prečo je X?“, kde X nemusí ani *dávať zmysel*, a už vôbec to nemusí byť tak.

„Prečo si *myslím*, že mám slobodnú vôľu?“ je v kontraste s tým zaručene zodpovedateľné. Je to fakt, že si myslíte, že máte slobodnú vôľu. Tento názor vyzerá omnoho pevnejší a uchopiteľnejší než prchavosť slobodnej vôle. A *naozaj* existuje nejaká pevná reťaz kognitívnych príčin a následkov, ktorá vedie k tomuto názoru.

Ak ste už vyrástli zo slobodnej vôle, vyberte si jednu z týchto náhradných otázok:

- „Prečo čas ide dopredu a nie dozadu?“ verzus „Prečo si myslím, že čas ide dopredu a nie dozadu?“
- „Prečo som sa narodil ako ja a nie ako niekto iný?“ verzus „Prečo si myslím, že som sa narodil ako ja a nie ako niekto iný?“
- „Prečo mám vedomie?“ verzus „Prečo si myslím, že mám vedomie?“
- „Prečo existuje skutočnosť?“ verzus „Prečo si myslím, že existuje skutočnosť?“

Krása tejto metódy je, že funguje bez ohľadu na to, či otázka *je alebo nie je* zmätená. Ako píšem tento text, mám na nohách ponožky. Mohol by som sa opýtať: „Prečo mám na nohách ponožky?“ alebo „Prečo si myslím, že mám na nohách ponožky?“ Opýtajme sa to druhé. Sledujúc naspäť reťaz kauzality, nachádzam:

- Myslím si, že mám ponožky, pretože vidím, že ich mám na nohách.

→ http://lesswrong.com/lw/og/wrong_questions/

- Vidím ponožky na svojich nohách, pretože na moju sietnicu dopadá svetlo v ponožkovom tvare.
- Svetlo v ponožkovom tvare dopadá na moju sietnicu, pretože sa odráža od ponožiek, ktoré mám na nohách.
- Odráža sa od ponožiek, ktoré mám na nohách, pretože na nohách mám ponožky.
- Na nohách mám ponožky, pretože som si ich obliekol.
- Obliekol som si ponožky, pretože som si myslel, že inak by mi bola zima na nohy.
- atď.

Hľadajúc späť v reťazi kauzality, krok za krokom, som objavil, že môj názor, že nosím ponožky, sa dá celkom vysvetliť faktom, že nosím ponožky. To je správne a primerané, pretože nemôžete o niečom získať informáciu bez toho, aby ste s tým interagovali.

Na druhej strane, ak vidím fatamorgánu jazera v púšti, správne kauzálne vysvetlenie mojej predstavy nezahŕňa fakt, že v púšti je nejaké skutočné jazero. V tomto prípade nie je moja viera v jazero iba *vysvetlená*, ale je *vyvrátená*.

V oboch prípadoch je však samotný názor skutočným javom, ktorý sa odohráva v skutočnom vesmíre – psychologické udalosti sú udalosti – a jeho kauzálna história sa dá vyhľadať.

„Prečo je uprostred púšte jazero?“ môže zlyhať, ak tam žiadne vysvetliteľné jazero nie je. Ale „Prečo *sa mi zdá*, že uprostred púšte je jazero?“ má vždy kauzálne vysvetlenie, také alebo onaké.

Možno tu niekto vidí príležitosť byť chytrý a povie: „Okej. Verím, že mám slobodnú vôľu, pretože mám slobodnú vôľu. Tak, hotovo.“ Samozrejme to nie je také jednoduché.

Môj vnem ponožiek na mojich nohách je udalosťou v zrakovej kôre. Fungovanie zrakovej kôry možno skúmať pomocou kognitívnej vedy, keby sa ukázalo mätúce.

Moja sietnica zachytáva svetlo nie pomocou mystickej procedúry vnímania, čarovného detektora ponožiek, ktorý sa z nevysvetliteľných dôvodov rozsvieti v prítomnosti ponožiek; je to mechanizmus, ktorý sa dá pochopiť pomocou biologických pojmov. Fotóny vstupujúce na sietnicu sa dajú pochopiť pomocou optických pojmov. Odrážavý povrch ponožky sa dá pochopiť pomocou elektromagnetických a chemických pojmov. Chlad na nohách sa dá pochopiť pomocou termodynamických pojmov.

Nie je to teda také jednoduché ako povedať: „Verím, že mám slobodnú vôľu, pretože ju mám – tak, hotovo!“ Musíte vedieť rozdeliť kauzálnu reťaz na menšie kroky a vysvetliť tieto kroky pomocou prvkov, ktoré samotné nie sú mätúce.

Mechanická interakcia mojej sietnice s mojimi ponožkami je dosť jasná, a dá sa opísať pomocou nemätúcich zložiek ako sú fotóny a elektróny. Kde je vo vašom mozgu detektor slobodnej vôle, a akým spôsobom detekuje prítomnosť alebo neprítomnosť slobodnej vôle? Ako tento senzor interaguje s vnímanou udalosťou, a čo sú mechanické detaily tejto interakcie?

Ak vaše presvedčenie vychádza z platného pozorovania skutočného javu, časom sa dostaneme k tomuto javu, ak začneme hľadať v kauzálnej reťazi od vášho názoru nazad.

Ak to, čo naozaj vidíte, je váš vlastný zmätok, hľadanie naspäť v reťazi kauzality nájde algoritmus, ktorý skresľuje skutočnosť.

V každom prípade táto otázka zaručene má nejakú odpoveď. Máte dokonca aj pekné konkrétne miesto, na ktorom začnete hľadať – svoj názor, pevne umiestnený vo vašej mysli.

Kognitívna veda možno nevyzerá taká vznešená a nádherná ako metafyzika. Ale otázky kognitívnej vedy sú aspoň *riešiteľné*. Nájsť odpoveď nemusí byť *ľahké*, ale odpoveď aspoň *existuje*.

Aha, a ešte: predstava, že kognitívna veda nie je taká vznešená a nádherná ako metafyzika, je jednoducho nesprávna. Niektorí čitatelia si to už začínajú všímať, aspoň dúfam.

*→

→ http://lesswrong.com/lw/oh/righting_a_wrong_question/

192. Klam projekcie mysle

V časoch úsvitu science fiction mimozemskí votrelci občas uniesli dievčinu v roztrhaných šatoch a odniesli ju s úmyslom zneuctiť, ako sa s láskou zobrazovalo na mnohých obálkach dávnych časopisov. Čo je čudné, mimozemšťania nikdy nešli po mužoch s roztrhanými tričkami.

Prečo by nehumaniodný mimozemšťan s odlišnou evolučnou históriou a evolučnou psychológiou sexuálne túžil po ľudskej žene? Vyzerá to nepravdepodobne. Jemne povedané.

Ľudia nerobia takéto chyby pri vedomom rozmýšľaní: „Všetky možné mysle sú pravdepodobne zostavené prakticky rovnakým spôsobom, preto by príšere s hmyzími očami ľudská samica pripadala príťažlivá.“ Umelec si pravdepodobne ani len nepoložil otázku, či mimozemšťan *vníma* ľudskú samicu ako príťažlivú. Namiesto toho, ľudská samica v roztrhaných šatoch *je sexi* – je to jej prirodzená vlastnosť.

Tí, ktorí sa pomýlili, nerozmýšľali nad mimozemšťanovou evolučnou históriou; sústredili sa na roztrhané šaty danej ženy. Keby tie šaty neboli roztrhané, tá žena by bola menej sexi; mimozemská príšera s tým nijako nesúvisí.

Zdá sa, že inštinktívne vnímame príťažlivosť ako priamu vlastnosť objektu Žena; Žena.príťažlivosť, podobne ako Žena.výška alebo Žena.hmotnosť.

Keby vás mozog používal takúto dátovú štruktúru, alebo niečo metaforicky podobné, potom by mu zvnútra pripadalo, že príťažlivosť je prirodzená vlastnosť danej ženy, nie vlastnosť mimozemšťana pozerajúceho na danú ženu. Keďže daná žena *je príťažlivá*, mimozemskej príšere sa bude páčiť – nie je to logické?

E. T. Jaynes používal pojem klam projekcie mysle, aby označil chybu premietania vlastností svojej vlastnej mysle do vonkajšieho sveta. Jaynes, neskorý veľmajster bayesovskej konšpirácie, sa najviac zaujímal o nesprávne zaobchádzanie s *pravdepodobnosťami* ako s vnútornými vlastnosťami predmetov, namiesto ako so stavmi čiastočného poznania v nejakej konkrétnej mysli. O chvíľu o tom poviem viac.

Lenže klam projekcie mysle je všeobecná chyba. Nachádza sa aj v hádke o skutočnom význame slova zvuk, aj na obálke časopisu s príšerou, ktorá odnáša ženu v roztrhaných šatoch, aj v Kantovom vyhlásení, že priestor je zo svojej samotnej podstaty plochý, aj v Humeovej definícii apriórnych myšlienok ako tých, ktoré „možno objaviť púhou myšlienkovou operáciou, nezávisle na tom, čo kde existuje vo vesmíre“...

(Zhodou okolností som raz čítal sci-fi príbeh o mužovi, ktorý nadviazal sexuálny vzťah s mysliacou mimozemskou rastlinou s vhodne tvarovanými listami; zistil, že je to samčia rastlina; chvíľu sa tým trápil; a nakoniec sa rozhodol, že na tom naozaj nezáleží. A vo Fogliovej a Pollottovej knihe *Nelegálni votrelci* ľudia pristanú na planéte obývanej mysliacim hmyzom, a vidia televíznu reklamu zobrazujúcu človeka, ako unáša chrobáka v jemných šifónových šatoch. Len som si myslel, že je to hodno spomenúť.)

*→

→ http://lesswrong.com/lw/oi/mind_projection_fallacy/

193. Pravdepodobnosť je v mysli

V predchádzajúcej kapitole som hovoril o klame projekcie mysle a dal som príklad mimozemskej príšery, ktorá unáša dievčinu v roztrhaných šatoch s úmyslom zneucti ju – chyba, ktorú som pripísal umelcovmu sklonu myslieť si, že príťažlivosť ženy je vlastnosťou samotnej ženy, Žena.príťažlivosť, a nie niečo, čo existuje v mysli pozorovateľa a pravdepodobne by to v mimozemskej mysli nebolo.

Pojem „klam projekcie mysle“ zaviedol veľký neskorý bayesovský majster, E. T. Jaynes, ako časť jeho dlhého a náročného boja proti zlorečeným frekventistom. Jaynes si myslel, že pravdepodobnosti sú v mysli, nie v prostredí – že pravdepodobnosti vyjadrujú nevedomosť, stav čiastočnej informácie; a keď neviem o nejakom jave, je to fakt o mojom stave mysle, nie fakt o danom jave.

Nedokážem túto pradávnu vojnu verne opísať niekoľkými slovami – ale klasický príklad argumentu vyzerá takto:

Máte mincu.

Tá minca je nevyvážená.

Vy neviete, akým spôsobom je nevyvážená, ani nakoľko je nevyvážená. Niekto vám akurát povedal: „Táto minca je nevyvážená“ a to je všetko, čo povedal.

To je celá a jediná informácia, ktorú máte.

Vyberiete mincu, vyhodíte ju a priplesnete rukou.

Teraz – skôr než zdvihnete ruku a pozriete sa na výsledok – ste ochotní povedať, že pripisujete pravdepodobnosť 0,5 tomu, že na minci padla hlava?

Frekventista povie: „Nie. Povedať ‚pravdepodobnosť 0,5‘ znamená, že táto minca má vnútorný sklon padať ako hlava rovnako často ako znak, takže keby sme túto mincu vyhodili nekonečne veľakrát, pomer hláv a znakov by sa blížil k 1 : 1. Lenže my vieme, že minca je nevyvážená, takže môže mať ľubovoľnú pravdepodobnosť padnutia ako hlava *okrem* 0,5.“

Bayesovec povie: „Neistota existuje na mape, nie v území. V skutočnom svete na tejto minci buď padla hlava, alebo na nej padol znak. Akékoľvek reči o ‚pravdepodobnosti‘ musia hovoriť o *informácii*, ktorú mám o danej minci – o mojom stave čiastočnej nevedomosti a čiastočnej vedomosti – nie iba o samotnej minci. Ďalej, máme všelijaké vety, ktoré ukazujú, že ak nebudem so svojím čiastočným poznaním zaobchádzať určitým spôsobom, budem uzatvárať hlúpe stávky. Ak musím plánovať, budem plánovať pre stav neistoty 50/50, kde nebudem výsledkom založeným na predpoklade, že padne hlava, pripisovať vyššiu váhu než výsledkom založeným na predpoklade, že padne znak. Môžete toto číslo nazvať, ako len chcete, ale musí sa riadiť zákonmi pravdepodobnosti, pod hrozbou hlúposti. Nemám teda najmenšie zábrany nazývať toto zvažovanie výsledku slovom pravdepodobnosť.“

Ja som na strane Bayesovcov. Možno ste si to o mne všimli.

Dokonca aj pred tým, ako hodíte vyváženú mincu, môže byť predstava o jej *vnútornej* pravdepodobnosti 50 % dopadnúť ako hlava jasne pomýlená. Možno tú minci držíte takým spôsobom, že je takmer isté, že na nej padne hlava, prípadne znak, podľa sily, akou ju hodíte, a vzdušných prúdov okolo vás. Ale ak neviete, akým spôsobom je minca nevyvážená, čo z toho?

Mám dojem, že bol súdny spor, kde niekto tvrdil, že povolávacia lotéria nebola férová, pretože lístky s menami neboli dostatočne dôkladne zamiešané; a sudca reagoval: „Voči komu bola neférová?“

Aby bol pokus s hádzaním mince opakovateľný, ako majú frekventisti sklon požadovať, mohli by sme postaviť automat na hádzanie mince, a overiť, že výsledky boli na 50 % hlava a na 50 % znak. Ale možno keby robot s extra citlivými očami a dobrými znalosťami fyziky sledoval automatický vrhač pripravený na hodenie, mohol by predpovedať dopadnutie mince – nie naisto, ale s pravdepodobnosťou 90 %. Aká by potom bola *skutočná* pravdepodobnosť?

Neexistuje „skutočná pravdepodobnosť“. Robot má jeden stav čiastočnej informácie. Vy máte iný stav čiastočnej informácie. Samotná minca nemá myseľ, a nepripisuje pravdepodobnosť ničomu; iba letí vzduchom, párkrát sa otočí, odrazí niekoľko molekúl vzduchu, a dopadne buď ako hlava alebo ako znak.

Toto je teda bayesovský pohľad na veci, a rád by som ukázal niekoľko klasických hlavolamov, ktorých *hlavolamovitosť* je odvodená od sklonu myslieť na pravdepodobnosti ako na vnútorné vlastnosti predmetov.

Vezmime si starú klasiku: Stretnete matematičku na ulici a ona spomenie, že sa jej postupne narodili dve deti. Opýtate sa: „Je aspoň jedno z tvojich detí chlapec?“ Matematička povie: „Áno, je.“

Aká je pravdepodobnosť, že má dvoch chlapcov? Ak predpokladáte, že pôvodná pravdepodobnosť dieťaťa byť chlapcom je ½, potom pravdepodobnosť, že má dvoch chlapcov pri danej informácii je 1/3. Pôvodné pravdepodobnosti boli: ¼ dvaja chlapci, ½ chlapec a dievča, ¼ dve dievčatá. Matematičkina odpoveď „áno“ má pravdepodobnosť ~1 v prvých dvoch prípadoch a ~0 v treťom. Renormalizácia nám necháva pravdepodobnosť 1/3 pre dvoch chlapcov a pravdepodobnosť 2/3 pre chlapca a dievča.

Ale predpokladajme, že by ste sa namiesto toho opýtali: „Je tvoje staršie dieťa chlapec?“ a matematička by bola odpovedala: „Áno.“ Potom by pravdepodobnosť, že matematička má dvoch chlapcov bola ½. Keďže staršie dieťa je chlapec, a mladšie dieťa môže byť čokoľvek.

Podobne, keby ste sa opýtali: „Je tvoje mladšie dieťa chlapec?“ Pravdepodobnosť, že sú obaja chlapci, by opäť bola ½.

Lenže, ak je aspoň jedna dieťa chlapec, musí buď platiť, že staršie dieťa je chlapec, alebo že mladšie dieťa je chlapec. Ako sa teda odpoveď v prvom prípade môže odlišovať od odpovedí na tie druhé?

Alebo tu je veľmi podobný problém: Povedzme, že mám štyri karty: srdcové eso, pikové eso, srdcovú dvojku a pikovú dvojku. Vyberiem dve karty náhodne. Opýtate sa ma: „Máš aspoň jedno eso?“ a ja odpoviem: „Áno.“ Aká je pravdepodobnosť, že držím dve esá? Je to 1/5. Existuje šesť možných kombinácií dvoch kariet, s rovnakou pôvodnou pravdepodobnosťou, a vy ste odstránili možnosť, že držím dve dvojky. Z piatich zostávajúcich kombinácií, iba jedna kombinácia obsahuje dve esá. Preto 1/5.

Predpokladajme, že by ste sa ma namiesto toho opýtali: „Máš pikové eso?“ Ak odpoviem „Áno,“ pravdepodobnosť, že tá druhá karta je srdcové eso, je 1/3. (Viete, že držím pikové eso, a existujú tri možnosti pre tú druhú kartu, z ktorých iba jedna je srdcové eso.) Podobne, ak sa ma opýtate: „Držíš srdcové eso?“ a ja odpoviem: „Áno,“ pravdepodobnosť, že držím pár es je 1/3.

Ale ako je možné, že ak sa ma opýtate: „Máš aspoň jedno eso?“ a ja poviem „Áno,“ pravdepodobnosť, že mám pár je 1/5? Buď musím držať pikové eso alebo srdcové eso, ako viete; a tak či onak je pravdepodobnosť, že držím pár es 1/3.

Ako je to možné? Zle som spočítal jednu alebo viac z týchto pravdepodobností?

Ak na to chcete prísť sami, urobte to teraz, lebo idem odhaliť...

Že všetky uvedené výpočty sú správne.

Čo sa paradoxu týka, žiaden tam nie je. *Dojem* pradoxu vychádza z myslenia, že pravdepodobnosti musia byť vlastnosťami samotných kariet. Eso, ktoré držím, musí byť buď srdcové alebo pikové; ale to neznamená, že vaša *vedomosť o* mojich kartách musí byť rovnaká, ako keby ste *vedeli*, že držím srdcové eso, alebo *vedeli*, že držím pikové eso.

Možno pomôže pomyslieť na Bayesovu vetu:

$P(H \mid E) = P(E \mid H) \times P(H) / P(E)$

Posledný člen, kde delíte výrazom P(E), je časť, v ktorej odhodíte všetky pravdepodobnosti, ktoré boli odstránené, a renormalizujete vaše pravdepodobnosti toho, čo zostáva.

Povedzme, že sa ma opýtate: „Máš aspoň jedno eso?“ *Predtým* než odpoviem, vaša pravdepodobnosť, že poviem „Áno“ by mala byť 5/6.

Ale ak sa ma opýtate: „Máš pikové eso?“, vaša pôvodná pravdepodobnosť, že poviem „Áno“, je iba ½.

Takže hneď vidíte, že ste sa v oboch prípadoch *dozvedeli* niečo veľmi odlišné. Budete odstraňovať iné možnosti, a renormalizovať pomocou iného P(E). Ak sa dozviete dve rôzne indície, nemali by ste byť prekvapení, ak skončíte s dvoma rôznymi stavmi čiastočnej informácie.

Podobne, ak sa opýtam matematičky: „Je aspoň jedno z tvojich dvoch detí chlapec?“, očakávam odpoveď „Áno“ s pravdepodobnosťou ¾, ale ak sa opýtam „Je tvoje staršie dieťa chlapec?“, očakávam

„Áno“ s pravdepodobnosťou ½. Nemalo by byť teda prekvapením, že skončím v rôznych stavoch čiastočnej vedomosti podľa toho, ktorú z týchto dvoch otázok som položil.

Jediný dôvod vidieť „paradox“ je myslenie, že pravdepodobnosť držania páru es je *vlastnosť kariet*, ktoré majú aspoň jedno eso, alebo vlastnosť *kariet*, ktoré majú zhodou okolností pikové eso. V tom prípade by bolo paradoxom, keby množiny kariet obsahujúce aspoň jedno eso mali vnútornú pravdepodobnosť páru 1/5, zatiaľ čo množiny kariet obsahujúce pikové eso by mali vnútornú pravdepodobnosť 1/3, a množiny kariet obsahujúcich srdcové eso by mali vnútornú pravdepodobnosť 1/3.

Podobne, ak si myslíte, že pravdepodobnosť 1/3 dvoch chlapcov je *vnútornou vlastnosťou* množiny detí, ktoré obsahujú aspoň jedného chlapca, potom to nie je konzistentné s tým, že množiny detí, z ktorých staršie je chlapec, majú *vnútornú* pravdepodobnosť ½, že to budú dvaja chlapci, a množiny detí, z ktorých mladšie je chlapec, majú vnútornú pravdepodobnosť ½, že to budú dvaja chlapci. To by bolo ako povedať: „Všetky zelené jablká vážia kilo, a všetky červené jablká vážia kilo, a všetky jablká, ktoré sú zelené alebo červené, vážia pol kila.“

Toto sa stane, ak začnete rozmýšľať, akoby pravdepodobnosti boli *vo* veciach, a nie že pravdepodobnosti sú stavy čiastočnej informácie *o* veciach.

Pravdepodobnosti vyjadrujú neistotu, a iba činitelia si môžu byť neistí. Prázdna mapa nezodpovedá prázdnemu územiu. Nevedomosť je v mysli.

*→

194. Citát nie je referent

V klasickej logike je operačná definícia totožnosti, že kedykoľvek platí veta „A = B“, môžete nahradiť „A“ za „B“ v hocijakej vete, kde sa objavuje B. Napríklad ak (2 + 2) = 4 je veta, a ((2 + 2) + 3) = 7 je veta, potom aj (4 + 3) = 7 je veta.

Toto vedie k problému, ktorý sa zvyčajne formuluje nasledujúcimi slovami: Zornička a večernica sú zhodou okolností ten istý predmet, planéta Venuša. Predpokladajme, že John vie, že zornička a večernica sú ten istý predmet. Mary si však myslí, že zornička je boh Lucifer, ale večernica je bohyňa Venuša. John si myslí, že Mary si myslí, že zornička je Lucifer. Musí si teda John (podľa substitúcie) myslieť, že Mary si myslí, že večernica je Lucifer?

Alebo tu je ešte jednoduchšia verzia problému. 2 + 2 = 4 je pravda; je to veta (((2 + 2) = 4) = PRAVDA). Veľká Fermatova veta je tiež pravda. Takže: Verím, že 2 + 2 = 4 => Verím, že PRAVDA => Verím, že Veľká Fermatova veta.

Áno, toto vyzerá *očividne* nesprávne. Ale predstavte si, že niekto píše logický uvažujúci program s použitím princípu „rovnaké pojmy možno vždy nahradiť“ a stane sa mu toto. Teraz si predstavte, že dotyčný píše článok o tom, ako tomuto zabrániť. Teraz si predstavte, že niekto iný nesúhlasí s jeho riešením. Táto hádka pokračuje dodnes.

Osobne by som povedal, že John robí typovú chybu, ako keby sa pokúša odčítať 5 gramov od 20 metrov. „Zornička“ nie je rovnakého typu ako zornička, takže to nie je tá istá vec. Názory nie sú planéty.

zornička = večernica

„zornička“ ≠ „večernica“

Problém podľa mňa vychádza z chyby v nedodržaní typového rozlíšenia medzi názormi a vecami. Pôvodná chyba bola napísať UI, ktorá si ukladá názory Mary na „zorničku“ pomocou rovnakej reprezentácie ako svoje vlastné názory na zorničku.

Ak Mary verí, že „zornička“ je Lucifer, to neznamená, že Mary verí, že „večernica“ je Lucifer, pretože „zornička“ ≠ „večernica“. Celý paradox vychádza z chyby nepoužívania úvodzoviek na správnych miestach.

→ http://lesswrong.com/lw/oj/probability_is_in_the_mind/

Môžete si spomenúť, že to nie je prvýkrát, čo hovorím o dodržiavaní typovej disciplány – naposledy to bolo, keď som hovoril o chybe v mýlení si očakávaného úžitku a úžitku. Je neuveriteľne užitočné, keď sa človek prvýkrát učí fyziku, naučiť sa sledovať svoje jednotky – môže to vyzerať otravné, stále písať „cm“ a „kg“ a tak ďalej, dokiaľ si nevšimnete, že (a) vaša odpoveď vyzerá byť rádovo nesprávna a (b) je vyjadrená v sekundách na štvorcový gram.

Podobne, názory sú iné veci ako planéty. Prinajmenšom ak hovoríme o ľudských názoroch, teda: Názory žijú v mozgu, planéty žijú vo vesmíre. Názory vážia pár mikrogramov, planéty vážia omnoho viac. Planéty sú väčšie než názory... ale asi je to jasné.

Samotné umiestnenie úvodzoviek okolo „zorničky“ zrejme nestačí niektorým ľuďom zabrániť, aby si ju mýlili so zorničkou, vďaka vizuálnej podobnosti textu. Možno by sa typová disciplína ľahšie uplatňovala s viditeľne odlišným kódovaním:

> zornička = večernica
>
> 007a.006f.0072.006e.0069.010d.006b.0061 ≠ 0076.0065.010d.0065.0072.006e.0069.0063.0061

Štúdium matematickej logiky vám môže tiež pomôcť naučiť sa odlišovať citát a referent. V matematickej logike |- P (P je veta) a |- [] „P“ (je dokázateľné, že existuje zakódovaný dôkaz zakódovanej vety P v nejakom kódovom dôkazovom systéme) sú veľmi odlišné výroky. Ak odhodíte úroveň úvodzoviek v matematickej logike, je to ako odhodiť metrickú jednotku vo fyzike – môžete odvodiť viditeľne smiešne výsledky ako „Rýchlosť svetla má dĺžku 299 792 458 metrov.“

Alfred Tarski raz skúšal definovať význam slova „pravda“ pomocou nekonečnej množiny viet:

> („Sneh je biely“ je pravda) vtedy a iba vtedy, keď (sneh je biely)
>
> („Lasice sú zelené“ je pravda) vtedy a iba vtedy, keď (lasice sú zelené)
>
> …

Keď vety ako tieto začnú vyzerať zmysluplne, budete vedieť, že ste začali rozlišovať medzi zakódovanými vetami a stavmi vonkajšieho sveta.

Podobne je pojem pravda pomerne odlišný od pojmu *skutočnosť*. Povedať „pravda“ znamená *porovnať* názor so skutočnosťou. Samotná skutočnosť nepotrebuje byť porovnávaná so žiadnymi názormi, aby bola skutočná. Pamätajte si to, keď bude nabudúce niekto tvrdiť, že nič nie je pravda.

* →

195. Kvalitatívny zmätok

Tvrdím, že základnou príčinou zmätku pri rozlišovaní medzi slovami „názor“, „pravda“ a „skutočnosť“ je kvalitatívne rozmýšľanie o názoroch.

Vezmite si archetypálny postmoderný pokus byť chytrý:

> „Slnko obieha okolo Zeme“ je pravda pre lovca a zberača Hungu, ale „Zem obieha okolo Slnka“ je pravda pre astronóma Amaru! Rôzne spoločnosti majú rôzne pravdy!

Nie, rôzne spoločnosti majú rôzne *názory*. Názor je iného typu ako pravda; je to ako porovnávať jablká s pravdepodobnosťami.

> Aha, ale nie je rozdiel medzi tým, ako používate slovo „názor“ a ako používate slovo „pravda“! Či poviete: „Myslím si ‚sneh je biely‘“ alebo poviete: „‚Sneh je biely‘ je pravda“, vyjadrujete presne ten istý názor.

Nie, tieto dve vety znamenajú pomerne odlišné veci, vďaka čomu si dokážem *predstaviť* možnosť, že moje názory sú nesprávne.

→ http://lesswrong.com/lw/ok/the_quotation_is_not_the_referent/

> Och, tvrdíš, že si to vieš *predstaviť*, ale nikdy tomu nebudeš *veriť*. Ako povedal Wittgenstein: „Keby existovalo sloveso s významom ‚nepravdivo veriť‘, nemalo by prvú osobu v prítomnom čase.“

A presne toto myslím tým, keď ukazujem na kvalitatívne myslenie ako zdroj problému. Binárny rozdiel medzi myslením si a nemyslením si je mätúco podobný rozdielu medzi pravdou a nepravdou.

Použime namiesto toho kvantitatívne uvažovanie. Povedzme, že pripisujem pravdepodobnosť 70 % tvrdeniu, že sneh je biely. Z toho vyplýva, že si myslím, že je šanca zhruba 70 %, že sa veta „sneh je biely“ ukáže ako pravdivá. Ak veta „sneh je biely“ je pravdivá, je moje pridelenie pravdepodobnosti 70 % tomuto výroku tiež „pravdivé“? Nuž, je pravdivejšie než keby som tomu priradil pravdepodobnosť 60 %, ale nie také pravdivé, ako keby som priradil pravdepodobnosť 80 %.

Keď hovoríme o súlade medzi priradením pravdepodobnosti a skutočnosťou, lepšie slovo než „pravda“ by bolo „presnosť“. „Presnosť“ znie kvantitatívnejšie, ako keď lukostrelec strieľa šíp: nakoľko blízko váš odhad pravdepodobnosti trafil k stredu terča?

Aby som skrátil dlhý príbeh, ukazuje sa, že existuje veľmi prirodzený spôsob hodnotenia presnosti priradenia pravdepodobnosti, v porovnaní so skutočnosťou: vezmite skrátka logaritmus pravdepodobnosti priradenej skutočnému stavu vecí.

Čiže ak je sneh biely, môj názor „70 %: ‚sneh je biely‘“ získa skóre -0,51 bitu: log2(0,7) = -0,51.

Ale čo ak sneh nie je biely, ako som pripustil s pravdepodobnosťou 30 %? Ak „sneh je biely“ je nepravda, môj názor „30 %: ‚sneh nie je biely‘“ získa -1,73 bitu. Všimnite si, že -1,73 < -0,51, takže som dopadol horšie.

Nakoľko presné si myslím, že sú moje vlastné názory? Nuž, môj odhad skóre je 70 % × -0,51 + 30 % × -1,73 = -0,88 bitu. Ak je sneh biely, potom sú moje názory správnejšie než očakávam; a ak sneh nie je biely, moje názory sú menej presné než očakávam; ale v žiadnom prípade nebude môj názor *presne* taký presný ako v priemere očakávam.

Toto všetko by sme si nemali pomýliť s výrokom: „Pripisujem pravdepodobnosť 70 %, že ‚sneh je biely‘.“ *Tomuto* výroku môžem veriť aj s pravdepodobnosťou ~1 – môžem si byť celkom istý, že toto je naozaj môj názor. Ak je to tak, potom budem očakávať, že môj meta-názor „~1: ‚Pripisujem pravdepodobnosť 70 %, že „sneh je biely“.‘“ získa ~0 bitov presnosti, čo je najlepší možný výsledok.

Len preto, že si nie som istý ohľadom snehu, neznamená, že si nie som istý ohľadom mojich *citovaných pravdepodobnostných názorov*. Sneh je tam vonku, moje názory sú vo mne. Môžem si byť omnoho menej neistý ohľadom toho, aký neistý som ohľadom snehu, než som si neistý ohľadom snehu. (Hoci názory o názoroch nie sú vždy presné.)

Porovnajme túto pravdepodobnostnú situáciu s kvalitatívnym uvažovaním, kde skrátka verím, že sneh je biely, a verím, že verím, že sneh je biely, a verím, že „‚sneh je biely‘ je pravda“, a verím, že „môj názor ‚„sneh je biely“ je pravda‘ je správny“, atď. Keďže všetky použité pravdepodobnosti sú 1, je ľahké si ich popliesť.

Tieto pekné rozlíšenia kvantitatívneho uvažovania však môžu spôsobiť skrat, ak si začnete myslieť „‚„sneh je biely“ s pravdepodobnosťou 70 %‘ je *pravda*“, čo je typová chyba. Je pravdivým faktom o vás, že *veríte* „pravdepodobnosť 70 %: ‚sneh je biely‘“; ale to neznamená, že samotné *priradenie* pravdepodobnosti môže byť „pravda“. Tento názor má skóre buď -0,51 bitu alebo -1,73 bitu presnosti, podľa skutočného stavu skutočnosti.

Znalci rozoznajú „‚„sneh je biely“ s pravdepodobnosťou 70 %‘ je pravda“ ako chybu myslenia si, že pravdepodobnosti sú vnútorné vlastnosti vecí.

Zvnútra naše názory o svete vyzerajú ako samotný svet, a naše názory o našich názoroch vyzerajú ako názory. Keď vidíte svet, prežívate názor zvnútra. Keď si všimnete, že niečomu veríte, prežívate názor o názore zvnútra. Ak sú teda vaše vnútorné reprezentácie názoru, a názoru o názore, rôzne, potom máte menší sklon ich pomiešať a dopustiť sa klamu projekcie mysle – aspoň dúfam.

Keď rozmýšľate v pravdepodobnostiach, vaše názory, a vaše názory o vašich názoroch, sa snáď nebudú reprezentovať natoľko podobne, aby ste si pomýlili názor s presnosťou, alebo pomýlili presnosť

so skutočnosťou. Keď rozmýšľate v pravdepodobnostiach *o svete*, vaše názory budú reprezentované s pravdepodobnosťami ∈ (0, 1). Na rozdiel od pravdivostných hodnôt výrokov, ktoré sú v { pravda, nepravda }. A čo sa týka presnosti vašich pravdepodobnostných názorov, tie môžete reprezentovať v rozsahu (-∞, 0). Vaše pravdepodobnosti *ohľadom vašich názorov* budú zvyčajne extrémne. A samotné veci – nuž, tie sú skrátka len červené alebo modré alebo vážiace 20 kíl alebo čokoľvek.

Takto azda máme menší sklon pomýliť si mapu s územím.

Toto typové rozlíšenie nám tiež môže pomôcť zapamätať si, že *neistota* je stav mysle. Minca nemá *vnútornú* pravdepodobnosť 50 %, ktorou stranou dopadne. Minca nespracováva názory a nemá o sebe čiastočnú informáciu. Pri kvalitatívnom uvažovaní si môžete vytvoriť názor, ktorý veľmi priamočiaro zodpovedá minci, napríklad „Na tejto minci padne hlava.“ Tento názor bude pravdivý alebo nepravdivý *v závislosti* od mince, a bude jasná implikácia z pravdivosti alebo nepravdivosti názoru na hornú stranu mince.

Ale aj pri kvalitatívnom uvažovaní, povedať, že samotná *minca* je „pravdivá“ alebo „nepravdivá“ by bola hrubá typová chyba. Minca nie je názor, je to minca. Územie nie je mapa.

Ak minca nemôže byť pravdivá alebo nepravdivá, ako by vôbec mohla sama sebe pripísať pravdepodobnosť 50 %?

*→

196. Rozmýšľaj ako skutočnosť

Vždy, keď počujem niekoho opisovať kvantovú fyziku ako „divnú“ – vždy, keď počujem niekoho plakať nad tajomným účinkom pozorovania na pozorované, alebo nad čudesnou existenciou nelokálnych korelácií, alebo nad neuveriteľnou nemožnosťou poznania polohy a hybnosti zároveň – pomyslím si: *Tento človek nikdy nebude rozumieť fyzike bez ohľadu na to, koľko kníh prečíta.*

Skutočnosť tu bola dávno predtým, než ste sa ukázali vy. Neurážajte ju škaredými menami ako „čudesná“ alebo „neuveriteľná“. Tento vesmír šíril komplexné amplitúdy konfiguračným priestorom desať miliárd rokov predtým, než sa na Zemi zjavil život. Kvantová fyzika nie je „divná“. *Vy* ste divní. *Vy* máte absolútne čudesnú predstavu, že skutočnosť by sa mala skladať z malých biliardových guličiek odrážajúcich sa od seba, zatiaľ čo skutočnosť je úplne normálny oblak komplexných amplitúd v konfiguračnom priestore. To je *váš* problém, nie problém skutočnosti, a *vy* ste ten, kto by sa mal zmeniť.

Ľudské intuície vytvorila evolúcia, a evolúcia je lacný trik. Ten istý optimalizačný proces, ktorý postavil vašu sietnicu naopak a potom musel viesť optický kábel cez vaše zorné pole, navrhol aj váš vizuálny systém, aby spracovával stále objekty odrážajúce sa v troch priestorových rozmeroch, pretože sa to hodilo na sledovanie tigrov. Avšak „tiger“ je iba presakujúce povrchné zovšeobecnenie – tigre vznikali postupne v priebehu evolúcie, a nie sú navzájom absolútne rovnaké. Keď zídete na základnú úroveň, na tú úroveň, kde sú zákony stabilné, globálne a bez výnimiek, nie sú tam žiadne tigre. V skutočnosti tam nie sú žiadne stále objekty odrážajúce sa v troch priestorových rozmeroch. Vyrovnajte sa s tým.

Nazývať skutočnosť „divnou“ vás udržiava v názore, ktorý sa už ukázal ako mylný. Teória pravdepodobnosti nám hovorí, že prekvapenie je mierou zlej hypotézy; ak je model konzistentne hlúpy – konzistentne naráža na udalosti, ktorým pripisuje nízke pravdepodobnosti – potom je načase tento model zahodiť. Dobrý model vykresľuje skutočnosť ako *normálnu*, nie divnú; dobrý model pripisuje vysokú pravdepodobnosť tomu, čo je naozaj. Intuícia je iba iné meno pre model: zlé intuície sú skutočnosťou šokované, dobré intuície cítia, že skutočnosť je prirodzená. Potrebujete si prispôsobiť intuície, aby vám vesmír pripadal normálny. Potrebujete myslieť ako skutočnosť.

Tento cieľový stav sa nedá vynútiť násilím. Bolo by zbytočné predstierať, že vám kvantová fyzika pripadá prirodzená, keď v skutočnosti vám pripadá čudná. To by akurát znamenalo popierať svoj zmätok, nie byť menej zmätený. Ale takisto vás bude brzdiť, keď si budete stále myslieť: *Aké čudesné!* Míňať emocionálnu energiu na nedôverčivosť je plytvanie časom, ktorý ste mohli použiť na aktualizáciu.

→ http://lesswrong.com/lw/om/qualitatively_confused/

Opakovane vás to vrhá späť do rámca starého nesprávneho pohľadu. Živí to váš pocit oprávneného rozhorčenia nad tým, že sa vám skutočnosť opovažuje protirečiť.

Tento problém sa netýka iba fyziky. Už ste sa niekedy pristihli, ako hovoríte veci typu: „Jednoducho nerozumiem, ako fyzik s PhD môže veriť v astrológiu?“ Nuž, ak doslova *nerozumiete*, naznačuje to problém vo vašom modeli ľudskej psychológie. Možno ste *rozhorčený* – chcete vyjadriť silné morálne odsúdenie. Ale ak doslova *nerozumiete*, potom vám vaše rozhorčenie bráni zmieriť sa so skutočnosťou. Nemalo by byť ťažké predstaviť si, ako fyzik s PhD začal veriť v astrológiu. Ľudia kompartmentalizujú, a to stačí.

Ja sa snažím vyhýbať fráze: „Jednoducho nerozumiem, ako...“ pri vyjadrovaní rozhorčenia. Ak *naozaj* nerozumiem, potom je môj model prekvapený faktmi, a mal by som ho zahodiť a nájsť si lepší model.

Prekvapenie existuje na mape, nie na území. Neexistujú prekvapujúce fakty, iba modely, ktoré sú faktmi prekvapené. To isté sa týka označovania faktov slovami ako „čudesný“, „neuveriteľný“, „nečakaný“, „zvláštny“, „nenormálny“ alebo „divný“. Keď zbadáte, že vás pokúša takáto nálepka, môže byť rozumné overiť si, či tento údajný fakt je naozaj faktom. Ale ak kontrola tento fakt potvrdí, potom problém nie je vo fakte, ale vo vás.

*→

197. Chaotické prevrátenie

Nedávno som sa rozprával s pár priateľmi na tému hodinovej produktivity a udržiavania sily vôle – niečo, s čím som celý svoj život zápasil.

Viem sa vyhnúť utekaniu pred ťažkým problémom, keď ho vidím po prvýkrát (vytrvalosť na časovej škále sekúnd), a dokážem vydržať pri tom istom probléme celé roky; ale udržať sa pri práci v škále *hodín*, to je pre mňa neustály boj. Netreba dodávať, že som už prečítal kopy a kopy rád; a najväčšia pomoc, ktorú som z toho dostal, bolo uvedomenie si, že značná časť iných tvorivých profesionálov má rovnaký problém, a tiež si s ním nevie poradiť, bez ohľadu na to, ako rozumne všetky tie rady znejú.

„Čo robíš, keď nedokážeš pracovať?“ opýtali sa ma priatelia. (Konverzácia asi neprebiehala presne takto, toto je len voľne vyjadrená podstata.)

Odpovedal som, že si zvyčajne pozerám náhodné webové stránky, alebo pozriem krátke video.

„Nuž,“ povedali, „ak vieš, že nejakú dobu nebudeš môcť robiť, mohol by si si pozrieť nejaký film alebo niečo.“

„Na nešťastie,“ odpovedal som, „musím robiť niečo, čo sa dá rozdeliť na krátke časové úseky, ako surfovať po webe alebo pozerať si krátke videá, lebo sa mi hocikedy môže stať, že opäť môžem pracovať, a neviem predvídať, kedy...“

A vtedy som sa zastavil, lebo som práve dostal osvietenie.

Vždy som myslel na svoj pracovný cyklus ako na niečo *chaotické*, niečo *nepredvídateľné*. Nikdy som tieto slová nepoužil, ale presne takto som s tým *zaobchádzal*.

Lenže tu moji priatelia naznačovali – aká čudná myšlienka – že *iní* ľudia by mohli predpovedať, kedy budú opäť schopní pracovať, a zariadiť si svoj čas podľa toho.

A po prvýkrát mi napadlo, že sa možno dopúšťam toho prekliateho starého hrdzavého Klamu projekcie mysle, priamo tu vo svojom každodennom živote, nie vo vysokej abstrakcii.

Možno moja produktivita nebola *nezvyčajne chaotická*; možno som iba *nezvyčajne hlúpy* ohľadom jej predpovedania.

Takto vyzerá prevrátená hlúposť – ako chaos. Niečo, čo sa ťažko zvláda, ťažko chápe, ťažko háda, niečo, s čím neviete urobiť nič. Nie je to iba poučka pre vysoko abstraktné veci ako je Umelá Inteligencia. Môže sa to týkať aj bežného života.

→ http://lesswrong.com/lw/hs/think_like_reality/

A dôvod, prečo nepomyslíme na alternatívne vysvetlenie „som hlúpy“, *nie je* – myslím si – že máme o sebe takú vysokú mienku. Je to skôr o tom, že na seba vôbec nemyslíme. Vidíme iba chaotické vlastnosti nášho prostredia.

Teraz mi teda napadlo, že problémom mojej produktivity nemusí byť chaos, ale moja vlastná hlúposť.

A to možno v niečom pomôže, a možno nepomôže. Iste to nevyriešilo celý problém ihneď. Povedať „neviem“ z vás neurobí znalca.

Ale je to prinajmenšom iná cesta než povedať „je to príliš chaotické“.

*→

198. Redukcionizmus

Takmer pred rokom, v apríli 2007, Matthew C napísal nasledujúce odporučenie témy pre *Overcoming Bias*:

> „Ako a prečo je súčasný hlavný filozofický pohľad (redukcionistický materializmus) samozrejme pravdivý […], zatiaľ čo hlavné filozofické pohľady všetkých minulých spoločností a civilizácií sú samozrejme podozrivé...“

Pamätám si to, lebo som sa pozrel na požiadavku a považoval som ju za legitímnu, ale nemohol som sa do tej témy pustiť, kým som nezačal postupnosť článkov o klame projekcie mysle, čo hodnú chvíľu trvalo...

Ale teraz je čas začať sa venovať tejto otázke. A hoci som ešte neprišiel k téme „materializmu“, môžeme začať „redukcionizmom“.

Po prvé, povedzme si na rovinu, že považujem „redukcionizmus“ v tom zmysle, ako toto slovo používam, za samozrejme správny; a nech ide do čerta každá minulá civilizácia, ktorá nesúhlasila.

Toto vyzerá ako veľmi silné tvrdenie, prinajmenšom jeho prvá časť. Všeobecná relativita vyzerá dobre potvrdená, ale ktovie, či ju nejaký budúci fyzik nevyvráti?

Na druhej strane, *späť* k newtonovskej mechanike sa nevrátime nikdy. Západka vedy sa otáča, ale neotáča sa naspäť. Existujú prípady v histórii vedy, kde teórie utrpela zranenie alebo dve a potom sa odrazila naspäť; ale keď nejaká teória dostane toľko šípov do hrude ako newtonovská mechanika, tak *zostane mŕtva*.

„Do čerta s tým, čo si minulé civilizácie mysleli“ vyzerá dosť bezpečne, keď minulé civilizácie verili v niečo, čo bolo falzifikované do odpadkového koša dejín.

A redukcionizmus nie je ani tak pozitívna hypotéza, ako *neprítomnosť* viery – konkrétne, neviera v istú podobu klamu projekcie mysle.

Raz som stretol známeho, ktorý tvrdil, že má skúsenosti ako námorný delostrelec a povedal: „Keď strieľaš delostrelecké granáty, musíš ich trasy počítať pomocou newtonovskej mechaniky. Keby si tie trasy počítal pomocou relativity, dostal by si nesprávny výsledok.“

A ja a ešte ďalší prítomný človek sme rovno povedali: „Nie.“ Ja som dodal: „Možno by si tie trasy nevedel spočítať dosť rýchlo, aby si tie odpovede dostal načas – asi to si tým myslel? Ale relativistický výsledok bude vždy presnejší než newtonovský.“

„Nie,“ povedal, „myslel som tým, že relativita ti dá *zlý výsledok*, pretože veci pohybujúce sa rýchlosťou delostreleckých granátov sa riadia newtonovskou mechanikou, nie relativitou.“

„Keby toto bola pravda,“ odpovedal som, „mohol by si to uverejniť vo fyzikálnom časopise a vyzdvihnúť si svoju Nobelovu cenu.“

Štandardná fyzika používa rovnakú *základnú* teóriu na opis letu lietadla Boeing 747, a na zrážky v relativistickom urýchľovači ťažkých iónov. Jadrá a lietadlá sa obidvoje, podľa nášho chápania, riadia špeciálnou relativiou, kvantovou mechanikou, a chromodynamikou.

→ http://lesswrong.com/lw/wb/chaotic_inversion/

My však používame celkom rôzne *modely* na chápanie aerodynamiky lietadla 747 a zrážok medzi jadrami zlata v urýchľovači. Počítačový model aerodynamiky lietadla 747 nemusí obsahovať jediný symbol, jediný bit v pamäti, ktorý by reprezentoval kvark.

Skladá sa teda lietadlo 747 z niečoho iného ako kvarkov? Nie, to iba vy ho *modelujete* pomocou *reprezentačných prvkov*, ktoré nezodpovedajú jedna k jednej kvarkom lietadla 747. Mapa nie je územie.

Prečo teda *nemodelovať* lietadlo 747 pomocou chromodynamickej reprezentácie? Pretože by vám trvalo gazilión rokov, kým by ste z modelu dostali nejakú odpoveď. A tiež by sa ten model nezmestil do pamäte všetkých počítačov na svete, v roku 2008.

Ako hovorí porekadlo: „Mapa nie je územie, ale územie si nemôžete poskladať a odložiť si ho do priehradky na rukavice." Niekedy potrebujete menšiu mapu, aby sa zmestila do tesnej priehradky na rukavice – ale to nemení samotné územie. Škála mapy nie je fakt o území, je to fakt o mape.

Keby *bolo* možné postaviť a spustiť chromodynamický model lietadla 747, dával by presné predpovede. Dokonca lepšie predpovede než aerodynamický model.

Aby sme postavili celkom presný model lietadla 747, nie je v princípe nutné, aby tento model obsahoval explicitné popisy vecí ako prúdenie vzduchu a zdvih. Nemusí existovať jediný symbol, jediný bit pamäte, ktorý zodpovedá polohe krídel. V princípe je možná postaviť presný model lietadla 747, ktorý nespomína nič *okrem* polí základných častíc a základných síl.

„Čože?" kričí antiredukcionista. „Hovoríš mi, že lietadlo 747 *v skutočnosti nemá krídla?* Ja tam tie krídla vidím!"

Pointa je rafinovaná. Nejde *len* o to, že predmet môže mať rôzne opisy na rôznych úrovniach.

Ide o to, že *samotné* „mať rôzne opisy na rôznych úrovniach" je niečo, čo hovoríme, čo správne patrí do oblasti hovorenia o mapách, a nie do oblasti hovorenia o území.

Nie je to tak, že by *samotné lietadlo, samotné zákony fyziky*, používali rôzne opisy na rôznych úrovniach – ako si myslel onen delostrelec. To *my*, pre naše pohodlie, používame rôzne zjednodušené modely na rôznych úrovniach.

Keby ste sa pozreli na definitívny chromodynamický model, taký čo by obsahoval iba polia elementárnych častíc a základných síl, tento model by obsahoval všetky fakty o prúdení vzduchu a zdvihu a polohe krídel – ale tieto fakty by boli implicitné, nie explicitné.

Vy, keby ste sa pozreli *na* tento model, a keby ste *o* tomto modeli rozmýšľali, by ste mohli zistiť, kde sú tie krídla. Keby ste to zistili, mali by ste vo svojej mysli explicitnú reprezentáciu polohy krídel – explicitný výpočtový objekt, vo vašej neurónovej pamäti. *Vo vašej mysli.*

Mohli by ste veru vydedukovať rôzne druhy explicitných popisov lietadla, na rôznych úrovniach, a dokonca aj explicitné pravidlá, ako vaše modely na rôznych úrovniach interagujú jeden s druhým, aby priniesli kombinované predpovede...

A algoritmu to zvnútra pripadá tak, že samotné lietadlo sa *zdá* byť zložené zároveň z mnohých úrovní, ktoré navzájom interagujú.

Názor *vyzerá zvnútra* tak, že sa zdá, že sa pozeráte priamo na skutočnosť. Keď sa *zdá*, že sa pozeráte na názor ako taký, vtedy v skutočnosti vnímate názor o názore.

Takže keď vaša myseľ zároveň verí explicitným popisom na mnohých rôznych úrovniach, a verí explicitným pravidlám na prechody medzi úrovňami, ako časť efektívneho kombinovaného modelu, *pripadá vám*, akoby ste videli systém, ktorý sa *skladá* z rôznych úrovní popisu a ich pravidiel na interakciu.

Ale toto sa iba mozog pokúša efektívne komprimovať predmet, ktorý nemá najmenšiu šancu začať modelovať na základnej úrovni. Lietadlo je príliš veľké. Dokonca aj atóm vodíka by bol príliš veľký. Interakcia kvarkov s kvarkmi sú šialene nezvládnuteľné. Nedokážete zvládnuť *pravdu*.

Ale fyzika funguje *naozaj* tak, aspoň pokiaľ vieme, že existuje *iba* jedna najzákladnejšia úroveň – polia elementárnych častíc a základné sily. Nedokážete zvládnuť holú pravdu, ale skutočnosť to dokáže zvládnuť bez najmenšieho zjednodušenia. (Rád by som vedel, odkiaľ Skutočnosť berie svoju výpočtovú kapacitu.)

Zákony fyziky neobsahujú osobitné dodatočné kauzálne prvky, ktoré by zodpovedali zdvihu alebo krídlam lietadla, spôsobom akým *myseľ inžiniera* obsahuje osobitné dodatočné *myšlienkové* prvky, ktoré zodpovedajú zdvihu alebo krídlam lietadla.

Toto, ako to ja vidím, je téza redukcionizmu. Redukcionizmus nie je pozitívna viera, ale skôr neviera v to, že vyššie úrovne zjednodušených viacúrovňových modelov naozaj sú tam vonku v území. Ak toto pochopíte na inštinktívne úrovni, rozpustí sa tým otázka: „Ako môžeš povedať, že lietadlo naozaj nemá krídla, keď ja tam tie krídla vidím?“ Tie kritické slová sú *naozaj* a *vidím*.

*→

199. Vysvetliť verzus vyvrátiť

Lamia od Johna Keatsa (1819)[192] si isto zaslúži nejakú odmenu za najslávnejšiu otravnú báseň:

> ...Neuletia snáď všetky kúzla
>
> Pri púhom dotyku chladnej filozofie?
>
> Kedysi bola v nebi úžasná dúha:
>
> Poznáme jej materiál, jej textúru; dali sme ju
>
> Do nudného katalógu bežných vecí.
>
> Filozofia strihá anjelom krídla,
>
> Dobýja všetky tajomstvá pravítkom a čiarami,
>
> Vyprázdni strašidelný vzduch, aj škriatkov z bane -
>
> Rozpára dúhu...

Moja zvyčajná odpoveď končí vetou: „Ak sa nedokážeme naučiť tešiť z púhej skutočnosti, naše životy budú naozaj prázdne.“ O tom poviem viac zajtra.

Dnes mám na mysli inú vec. Vezmime si len tieto riadky:

> Vyprázdni strašidelný vzduch, aj škriatkov z bane -
>
> Rozpára dúhu...

Zdá sa, že „púhy dotyk chladnej filozofie“, čiže pravda, zničil:

- Prízraky vo vzduchu
- Škriatkov v bani
- Dúhy

To mi pripomína celkom inú básničku:

> Jedna z týchto vecí
>
> Nie je ako ostatné
>
> Jedna z týchto vecí
>
> Sem nepatrí

Vzduch bol zbavený prízrakov, baňa bola zbavená škriatkov – ale dúha je stále tu!

V kapitole „Napravenie nesprávnej otázky“ som napísal:

> Hľadajúc späť v reťazi kauzality, krok za krokom, zisťujem, že moja viera, že nosím ponožky, je plne vysvetlená faktom, že nosím ponožky... Na druhej strane, ak vidím fatamorgánu jazera v púšti, správna kauzálne vysvetlenie mojej vízie nezahŕňa fakt, že v púšti

→ http://lesswrong.com/lw/on/reductionism/

192 John Keats, „Lamia,“ *The Poetical Works of John Keats* (London: Macmillan) (1884).

je nejaké skutočné jazero. V tomto prípade moja viera v jazero nebola iba *vysvetlená*, ale bola *vyvrátená*.

Dúha bola *vysvetlená*. Prízraky vo vzduchu a škriatkovia v bani boli *vyvrátení*.

Myslím si, že toto je kľúčové rozlíšenie, ktoré antiredukcionisti nechápu ohľadom redukcionizmu.

Nepochopenie tohto rozlíšenia môžete vidieť v klasickej námieste voči redukcionizmu:

> Ak je redukcionizmus správny, potom aj tvoja viera v redukcionizmus je iba výsledkom pohybu molekúl – prečo by som mal počúvať, čo hovoríš?

Kľúčové slovo v uvedenom je *iba*; slovo, ktoré naznačuje, že prijatie redukcionizmu by *vyvrátilo* všetky myšlienkové procesy, ktoré viedli k môjmu prijatiu redukcionizmu, rovnako ako sa *vyvráti* optický klam.

Ale vy môžete vysvetliť, ako funguje poznávací proces, bez toho, aby bol „iba"! Moja viera, že nosím ponožky je iba výsledkom toho, že moja zraková kôra rekonštruuje nervové impulzy poslané z mojej sietnice, ktorá zachytila fotóny odrazené z mojich ponožiek... čo znamená, že podľa vedeckého redukcionizmu je moja viera, že nosím ponožky, iba výsledkom faktu, že nosím ponožky.

Čo sa môže odohrávať v antiredukcionistických mysliach, keď dávajú dúhu a vieru v redukcionizmus do tej istej kategórie ako prízraky a škriatkov?

Deje sa tam niekoľko vecí naraz. Ale teraz sa sústreďme na základnú myšlienku uvedenú včera: Klam projekcie mysle medzi mnohoúrovňovou mapou a jednoúrovňovým územím.

(Čiže: Neexistuje spôsob, ako by ste vy mohli modelovať lietadlo 747 kvark za kvarkom, preto *musíte* použiť viacúrovňovú mapu s explicitnými poznávacími reprezentáciami krídel, prúdenia vzduchu, a tak ďalej. To neznamená, že existuje viacúrovňové územie. Skutočné zákony fyziky, podľa nášho najlepšieho poznania, sa týkajú iba polí elementárnych častíc.)

Myslím si, že keď fyzici povedia „Neexistujú žiadne *základné* dúhy," antiredukcionisti počujú: „Neexistujú žiadne dúhy."

Ak nerozlišujete medzi mnohoúrovňovou mapou a jednoúrovňovým územím, potom keď sa vám niekto pokúša vysvetliť, že dúha nie je základná vec vo fyzike, prijatie tohto vám *pripadá ako* vymazanie dúhy z vašej mnohoúrovňovej mapy, čo vám *pripadá ako* vymazanie dúhy zo sveta.

Keď Veda povie: „tigre nie sú *elementárne* častice, skladajú sa z kvarkov," antiredukcionista to počuje ako rovnaký typ odmietnutia ako: „pozreli sme sa, či je vo vašej garáži drak, ale bol tam iba prázdny vzduch."

Čo vedci urobili s dúhou a čo vedci urobili so škriatkami, zrejme pripadalo Keatsovi rovnaké...

Na podporu tejto pod-tézy som v diskusii ku Keatsovej básni úmyselne použil niekoľko vyjadrení, ktoré boli klamy projekcie mysle. Ak ste si to nevšimli, to zrejme naznačuje, že takéto klamy sú také časté, že ich človek prejde bez povšimnutia.

Napríklad:

> „Vzduch bol zbavený prízrakov, baňa bola zbavená škriatkov – ale dúha je stále tu!"

V skutočnosti Veda zbavila *model* vzduchu *viery* v prízraky, a zbavila *mapu* bane *reprezentácie* škriatkov. Veda nemohla v skutočnosti – ako ju obviňuje samotná Keatsova báseň – vziať krídla skutočnému anjelovi a zničiť ich chladným dotykom pravdy. V skutočnosti v tom vzduchu *nikdy neboli* žiadne prízraky, ani škriatkovia v bani.

Ďalší príklad:

> „Čo vedci urobili s dúhou a čo vedci urobili so škriatkami, zrejme pripadalo Keatsovi rovnaké."

Vedci nič *neurobili so* škriatkami, iba so „škriatkami". Citát nie je referent.

Ale ak podľahnete klamu projekcie mysle – a naše názory nám štandardne pripadajú jednoducho ako stav ako svet *je* – potom v čase T = 0 bane (údajne) obsahovali škriatkov; v čase T = 1 nejaký vedec pretancoval po scéne; a v čase T = 2 sú bane (údajne) prázdne. Je jasné, že tam kedysi boli škriatkovia, ale ten vedec ich zabil.

Zlý vedec! Nedostaneš žiadnu básničku, vrah škriatkov!

Nuž, takto vám to *pripadá*, ak sa emocionálne pripútate k škriatkom, a potom nejaký vedec povie, že škriatkovia neexistujú. Vyžaduje si to silnú myseľ, hlbokú úprimnosť, a vedomé úsilie povedať v tomto bode: „Čo môže byť zničené pravdou, by malo byť zničené“ a „Ten vedec neodstránil škriatkov, odstránil iba môj blud; nebol som pripravený o nič, čo by mi *právom* patrilo“ a „Ak škriatkovia existujú, chcem veriť, že škriatkovia existujú; ak škriatkovia neexistujú, chcem veriť, že škriatkovia neexistujú; nechcem byť pripútaný k názorom, ktoré možno nechcem mať“ a všetky ďalšie veci, ktoré si racionalisti majú v takýchto prípadoch hovoriť.

Ale čo sa týka dúhy, netreba ísť tak ďaleko. Tá dúha tam *stále je!*

*→

200. Falošný redukcionizmus

> Kedysi bola v nebi úžasná dúha:
>
> Poznáme jej materiál, jej textúru; dali sme ju
>
> Do nudného katalógu bežných vecí.
>
> --John Keats, Lamia

Tipujem – ale je to iba tip – že samotný Keats *nepoznal* materiál a textúru dúhy. Nie tak, ako Newton poznal dúhy. Dokonca možno vôbec. Možno Keats iba niekde čítal, že Netwon vysvetlil dúhu ako „svetlo odrazené od dažďových kvapiek“...

...čo bolo v skutočnosti známe v 13-tom storočí. Newton iba pridal upresnenie, keď ukázal, že svetlo sa rozdeľuje na farebné zložky, nie že by sa menila farba. Ale to vrátilo dúhy na titulné stránky novín. A tak si Keats spolu s Charlesom Lambom a Williamom Wordsworthom a Benjaminom Haydonom pripili na „hanbu Newtonovej pamiatky“, pretože „zničil poéziu dúhy tým, že ju zredukoval na hranol.“ To je jeden z dôvodov pochybovať, že Keats do hĺbky rozumel danej téme.

Tipujem, ale je to iba tip, že by Keats *nebol* dokázal načrtnúť na papier, prečo sa dúha objavuje iba keď je Slnko za vašou hlavou, alebo prečo je dúha oblúk kružnice.

Ak je to tak, Keats mal Falošné Vysvetlenie. V tomto prípade, *falošnú redukciu*. Bolo mu *povedané, že* dúha bola redukovaná, ale v skutočnosti *nebola redukovaná* v jeho modeli sveta.

Toto je ďalší z tých rozdielov, ktoré antiredukcionistom nedochádzajú – rozdiel medzi vyznaním holého faktu, že niečo je redukovateľné, a *videním* ako.

Za toto by sme nemali antiredukcionistov príliš obviňovať, pretože je to časť všeobecného problému.

Už som písal o zdanlivom poznaní, ktoré nie je poznanie, a názoroch, ktoré nie sú *o* ich domnelých objektoch, ale sú to iba záznamy na odrecitovanie v triede, a slová, ktoré fungujú ako zastavovače zvedavosti namiesto ako odpovede, a technoslangu, ktorý iba vyjadruje členstvo v literárnom žánri „vedy“...

Je veľmi veľký rozdiel medzi schopnosťou *vidieť*, odkiaľ pochádza dúha, a hraním sa s hranolmi, aby sa to potvrdilo, a možno urobením si vlastnej dúhy rozprašovaním vodných kvapiek...

...verzus keď vám nejaký odmeraný filozof iba *povie*: „Nie, na dúhe nie je nič zvláštne. Nepočuli ste? Vedci ju vyvrátili. Iba to súvisí s dažďovými kvapkami, alebo také niečo. Niet sa nad čím vzrušovať.“

Myslím si, že tento rozdiel pravdepodobne môže za pekelne veľa zo smrtiacej existenciálnej prázdnoty, ktorá údajne sprevádza vedecký redukcionizmus.

Musíte si preložiť antiredukcionistické skúsenosti s „redukcionizmom“ nie v zmysle, že *naozaj videli*, ako dúha funguje, ani nie v zmysle, že mali svoje kritické „Aha!“, ale v zmysle, že im bolo

→ http://lesswrong.com/lw/oo/explaining_vs_explaining_away/

povedané, že heslo je „Veda“. Výsledkom je iba presunutie dúh do iného *literárneho žánru* – literárneho žánru, ktorý sa naučili považovať za nudný.

Pre nich je počuť „Veda vysvetlila dúhy!“ ako zavesiť na dúhy značku: „Tento jav bol vyhlásený za NUDNÝ rozhodnutím Rady Sofistikovaných Literárnych Kritikov. Rozíďte sa.“

A to je všetko, čo tá značka hovorí: iba to a nič iné.

Literárni kritici teda prišli o svojich škriatkov násilím; neboli rozpustení vhľadom, ale odstránení na strohý príkaz autority. Nedostali žiadnu krásu výmenou za vzduch bez prízrakov, žiadne skutočné porozumenie, ktoré by mohlo byť samo osebe zaujímavé. Iba nálepku hovoriacu: „Ha! Myslel si si, že dúhy sú pekné? Hlúpy, nevzdelaný blázon. Toto je časť literárneho žánru vedy, kde sú suché a vážne nezrozumiteľné slová.“

Takto antiredukcionisti zažívajú „redukcionizmus“.

Nuž, nemôžem za to viniť Keatsa, chudák chlapec asi nebol správne vychovaný.

Ale opovážiť sa pripiť na „hanbu Newtonovej pamiatky?“

Navrhujem „na pamiatku Keatsovej hanby“ ako prípitok pre racionalistov. Na zdravie!

* →

201. Básnici zo savany

> Básnici hovoria, že veda berie krásu hviezdam – sú to iba hrudy atómov plynu. Nič nie je „iba“. Aj ja vidím hviezdy v nočnej púšti, a cítim ich. Ale vidím viac alebo menej?
>
> Rozľahlosť nebies naťahuje moju predstavivosť – moje malé oko zaseknuté na tomto kolotoči dokáže zachytiť milión rokov staré svetlo. Obrovský vzor – ktorého som časťou – možno moja hmota bola vyvrhnutá nejakou zabudnutou hviezdou, ako je jedna z tamtých. Alebo si ich pozrite väčším okom v Palomare, ako sa ponáhľajú z nejakého spoločného začiatočného bodu, kde azda boli všetky spolu. Čo je ten vzor, alebo význam, alebo prečo? Tomu tajomstvu neuškodí, ak o ňom trochu vieme.
>
> Pretože pravda je omnoho úžasnejšia než si akíkoľvek minulí umelci predstavovali! Prečo o tom nehovoria dnešní básnici?
>
> Akí ľudia sú básnikmi, čo dokážu hovoriť o Jupiterovi, keby bol ako človek, ale ak je to obrovská točiaca sa guľa metánu a amoniaku, musia mlčať?[193]
>
> --Richard Feynman, Feynmanove lekcie fyziky[193], Časť I, str. 3-6

Toto je skutočná otázka, tam na poslednom riadku – aký druh básnika dokáže písať o bohu Jupiterovi, ale nie o obrovskej guli Jupiteri? Či myslel Feynman túto otázku rečnícky alebo nie, má to skutočnú odpoveď:

Ak je Jupiter ako my, môže sa zamilovať, a stratiť lásku, a opäť ju získať.

Ak je Jupiter ako my, môže sa usilovať, a stúpať, a byť zvrhnutý.

Ak je Jupiter ako my, môže sa smiať alebo plakať alebo tancovať.

Ak je Jupiter obrovská rotujúca guľa z metánu a amoniaku, je pre básnika omnoho ťažšie vzbudiť v nás pocity.

Existujú básnici a rozprávkari, ktorí hovoria, že Veľké Príbehy sú nadčasové a nikdy sa nemenia, iba sa opakovane rozprávajú. Hovoria s pýchou, že Shakespeare a Sofokles sú zviazaní putom remesla silnejšieho než púhe stáročia; že títo dvaja dramatici sa mohli vymeniť v čase a nič by sa nezmenilo.

Donald Brown raz zostavil zoznam vyše dvesto „ľudských univerzálií“ nachádzajúcich sa v každej študovanej ľudskej kultúre (alebo v ich drvivej väčšine), od San Francisca po !Kungov v púšti Kalahari. Na zozname je manželstvo, a vyhýbanie sa incestu, a materinská láska, a súperenie súrodencov, a hudba a

→ http://lesswrong.com/lw/op/fake_reductionism/

193 Richard P. Feynman, Robert B. Leighton, and Matthew L. Sands, *The Feynman Lectures on Physics*, 3 vols. (Reading, MA: Addison-Wesley, 1963).

závisť a tanec a rozprávanie a estetika, a rituálna mágia na liečenie chorých, a básne v hovorených riadkoch oddelených pauzami...

Od nikoho, kto vie niečo o evolučnej psychológii, nemožno očakávať, že by to poprel: Tie najsilnejšie emócie máme hlboko vryté, v krvi a kostiach, v mozgu a DNA.

Možno by to chcelo trochu doladiť, ale asi by ste *mohli* porozprávať „Hamleta“ sediac pri ohnisku v pravekej savane.

Takže človek vidí, prečo John „párame dúhu“ Keats mohol mať pocit, že sa niečo stratilo, keď mu povedali, že dúha je slnečné svetlo odrazené od dažďových kvapiek. Dažďové kvapky netancujú.

V Starom Zákone sa píše, že Boh raz zničil svet potopou, ktorá prikryla celú súš, utopila všetkých strašne vinných mužov a ženy sveta spolu s ich strašne vinnými bábätkami, ale Noe postavil gigantickú drevenú archu, atď., a keď bola väčšina ľudského druhu vyhladená, Boh dal na oblohu dúhu ako znamenie, že už to druhýkrát neurobí. Prinajmenšom nie vodou.

Môžete vidieť, ako by bol Keats *šokovaný*, že moderná veda protirečí tomuto krásnemu príbehu. Najmä ak (ako som napísal v predchádzajúcej kapitole) Keats skutočne dúham nerozumel, nemal vhľad „Aha!“, ktorý by bolo byť fascinujúci sám osebe, aby nahradil odobratú drámu...

Ach, ale možno by Keats bol právom sklamaný *aj keby* poznal danú matematiku. Biblický príbeh o dúhe je príbehom krvilačnej vraždy a usmievavého šialenstva. Ako by mohlo čokoľvek o dažďových kvapkách a lome svetla primerane nahradiť toto? Dažďové kvapky nekričia, keď zomierajú.

Veta teda zobrala romantiku (hovorí básnik obdobia romantizmu) a čo vám dá naspäť, sa nikdy nevyrovná dráme originálu...

(to jest, pôvodnému bludu)

...dokonca ani keď poznáte rovnice, pretože tie rovnice nie sú o silných emóciách.

Toto je tá najsilnejšia odpoveď, akú si viem predstaviť, že by nejaký básnik obdobia romantizmu mohol povedať Feynmanovi – hoci si nepamätám, či som niekedy niečo také počul povedať.

Viete si domyslieť, že nesúhlasím s básnikmi obdobia romantizmu. Takže môj vlastný postoj je takýto:

Nie je *potrebné*, aby Jupiter bol ako človek, pretože *ľudia* sú ako ľudia. Ak je Jupiter obrovská rotujúca guľa z metánu a amoniaku, to neznamená, že z vesmíru zmizla láska a nenávisť. Vo vesmíre stále *sú* milujúce a nenávisiace mysle. *My*.

Keď je nás podľa posledného sčítania šesť miliárd, naozaj musí aj Jupiter byť na zozname možných protagonistov?

Nie je *nevyhnutné* hovoriť Veľké Príbehy o planétach alebo dúhach. Odohrávajú sa po celom svete, každý deň. Každý deň niekto niekoho zabije z pomsty; každý deň niekto omylom zabije priateľa; každý deň sa státisíce ľudí zamilujú. A aj keby to tak nebolo, mohli by ste písať fikcie o ľuďoch – nie o Jupiteri.

Zem je stará a odohrávala tie isté príbehy mnohokrát pod Slnkom. Rozmýšľam, či by nemohlo byť načase, aby sa niektoré z týchto Veľkých Príbehov zmenili. Pre mňa prinajmenšom príbeh zvaný „Zbohom“ stratil svoje čaro.

Veľké Príbehy nie sú nadčasové, pretože ľudský druh nie je nadčasový. Vráťte sa dostatočne ďaleko v evolúcii hominidov, a nikto nebude rozumieť *Hamletovi*. Vráťte sa dostatočne ďaleko v čase, a nenájdete žiadne mozgy.

Veľké Príbehy nie sú večné, pretože ľudský druh, *Homo sapiens sapiens* nie je večný. S najväčšou úprimnosťou pochybujem, že máme pred sebou ešte ďalších tisíc rokov v terajšej podobe. Nehovorím to so smútkom: Myslím si, že máme na viac.

Nerád by som videl, aby by sa všetky Veľké Príbehy celkom stratili, v budúcnosti. Vidím veľmi malý rozdiel medzi takýmto výsledkom, a pádom Slnka do čiernej diery.

Ale Veľké Príbehy v ich terajších podobách *už boli vyrozprávané*, znova a znova. Nemyslím to v zlom, ale niektoré z nich by mali zmeniť svoju podobu, alebo diverzifikovať svoje konce.

„A potom žili šťastne až naveky“ by sa oplatilo skúsiť aspoň raz.

Veľké Príbehy sa môžu diverzifikovať, a mali by sa, ako ľudstvo rastie. Časť tejto morálky je myšlienka, že keď nájdeme zvláštnosť, mali by sme si ju vážiť natoľko, že jej príbeh povieme pravdivo. Aj keď nám to trochu skomplikuje písanie poézie.

Ak ste dosť dobrý básnik, aby ste napísali ódu na obrovskú rotujúcu guľu z metánu a amoniaku, píšete niečo *originálne*, o novo objavenej časti skutočného sveta. Možno to nebude také dramatické alebo pútavé ako Hamlet. Ale príbeh o Hamletovi bol už povedaný! Ak píšete o Jupiterovi akoby to bol človek, potom svoju mapu vesmíru zbavujete kúska zložitosti; tlačíte Jupiter do formy príbehov, ktoré už boli povedané na Zemi.

Jamesova Thomsonova „Svätá báseň na pamiatku Sira Isaaca Newtona“, ktorá chváli dúhu za to, čím je naozaj – môžete sa hádať, či Thomsonova báseň je alebo nie je taká pútavá ako Johnova Keatsova Lamia, ktorá bola milovaná a stratená. Ale príbehy o láske a strate a cynizme *už boli povedané*, už dávno v starovekom Grécku, a nepochybne mnohokrát predtým. Dokiaľ sme nepochopili dúhu ako vec *odlišnú* od príbehov mágie ľudského tvaru, skutočný príbeh dúhy nemohol byť zveršovaný.

Hranica medzi vedeckou fantastikou a vesmírnou operou bola raz nakreslená takto: Ak môžete vziať zápletku príbehu a umiestniť ju do Divokého Západu alebo do stredoveku a nič sa nezmení, potom to nie je *skutočná* vedecká fantastika. V skutočnej vedeckej fantastike je veda podstatou časťou zápletky – nemôžete presunúť príbeh z vesmíru do savany a nič pritom nestratiť.

Richard Feynman sa pýtal: „Akí ľudia sú básnikmi, čo dokážu hovoriť o Jupiterovi, keby bol ako človek, ale ak je to obrovská točiaca sa guľa metánu a amoniaku, musia mlčať?“

Sú to *básnici zo savany*, ktorí dokážu hovoriť iba také príbehy, ktoré by dávali zmysel pri ohnisku pred desať tisíc rokmi. Básnici zo savany, ktorí dokážu hovoriť *iba* Veľké Príbehy v ich klasických podobách, a nič viac.

*→

→ http://lesswrong.com/lw/oq/savanna_poets/

Q: Radosť z púhej skutočnosti

202. Radosť z púhej skutočnosti

> ...Neuletia snáď všetky kúzla
>
> Pri púhom dotyku chladnej filozofie?
>
> Kedysi bola v nebi úžasná dúha:
>
> Poznáme jej materiál, jej textúru; dali sme ju
>
> Do nudného katalógu bežných vecí.
>
> --John Keats, Lamia

> „Nič nie je ‚iba‘.“
>
> --Richard Feynman

Musíte obdivovať tú frázu „nudný katalóg bežných vecí“. Čo presne je to, čo patrí do tohto katalógu? Myslím, okrem dúhy.

Nuž veci, ktoré sú obyčajné, samozrejme. Veci, ktoré sú normálne; veci, ktoré sú nemagické; veci, ktoré sú známe alebo poznateľné; veci, ktoré hrajú podľa pravidiel (podľa *hocijakých* pravidiel, lebo to ich robí nudnými); veci, ktoré sú časťou bežného vesmíru; veci, ktoré sú, jedným slovom, *skutočné*.

Tak tomuto hovorím vykopať si poriadnu jamu.

Takýmto tempom budete skôr či neskôr sklamaní zo *všetkého* – buď sa ukáže, že to neexistuje, alebo ešte horšie, ukáže sa, že je to naozaj.

Ak sa nedokážeme tešiť z vecí, ktoré sú iba skutočné, naše životy budú *vždy* prázdne.

Pre aký hriech boli dúhy degradované do nudného katalógu bežných vecí? Pre hriech, že mali vedecké vysvetlenie. „Poznáme jej materiál, jej textúru,“ hovorí Keats – zaujímavé použitie slova „my“, pretože si myslím, že samotný Keats nepoznal toto vysvetlenie. Myslím si, že už len počuť, že niekto iný vie, bolo naňho priveľa. Myslím si, že samotná predstava, že dúhy sú *v princípe* vedecky vysvetliteľné, bola preňho priveľa. A ak aj Keats takto nerozmýšľal, nuž, poznám veľa ľudí, ktorí áno.

Už som poznamenal, že nič nie je *vnútorne* tajomné – teda nič, čo naozaj existuje. Ak nerozumiem nejakému javu, je to fakt o mojom stave mysle, nie fakt o danom jave; uctievať jav, pretože vyzerá taký úžasne tajomný, je uctievať svoju vlastnú nevedomosť; prázdna mapa nezodpovedá prázdnemu územiu, je to iba miesto, ktoré sme ešte nenavštívili, atď. Atď...

Čo znamená, že *všetko* – všetko, čo naozaj existuje – má potenciál skôr či neskôr skončiť v „nudnom katalógu bežných vecí“.

Máte na výber buď:

- Rozhodnúť sa, že veci majú dovolené byť nemagické, poznateľné, vedecky vysvetliteľné, iným slovom *skutočné*, a stále nám na nich môže záležať;
- Alebo vás bude zvyšok vášho života trápiť existenciálna nuda, ktorá sa *nedá vyliečiť*.

(Sebaklam môže byť možnosťou pre druhých, ale nie pre vás.)

Toto dáva trochu odlišný pohľad na čudesný zvyk, ktorému sa oddávajú čudní ľudia zvaní *vedci*, keď ich odrazu začnú fascinovať zvyšky vlákniny alebo vtáči trus alebo dúhy, alebo nejaká iná obyčajná vec, ktorej by svetaskúsení a sofistikovaní ľudia nikdy nevenovali druhý pohľad.

Mohli by ste povedať, že vedci – prinajmenšom *niektorí* vedci – sú tí ľudia, ktorí sú *v princípe* schopní tešiť sa zo života v skutočnom vesmíre.

*→

→ http://lesswrong.com/lw/or/joy_in_the_merely_real/

203. Radosť z objavu

> „Newton bol najväčším géniom, aký kedy žil, a najšťastnejším; nemôžeme totiž zaviesť poriadok sveta viac než raz.“
>
> --Lagrange

Viac ma baví objavovať veci sám, než čítať o nich v učebniciach. To je správne a vhodné, a dalo sa očakávať.

Ale objaviť niečo, čo *nikto iný nevie* – byť *prvým*, ktorý odhalí tajomstvo...

Existuje príbeh, že jeden z prvých mužov, ktorý si uvedomili, že hviezdy horia fúziou – uveriteľne som to videl pripísané Fritzovi Houtermansovi a Hansovi Bethemu – sa v noci prechádzal s priateľkou, ona poznamenala, aké krásne sú hviezdy, a on odpovedal: „Áno, a práve teraz som jediný na svete, ktorý vie, prečo svietia.“

Početné zdroje potvrdujú, že tento zážitok, byť prvým človekom, ktorý vyrieši veľké tajomstvo, je *ohromne* povznášajúci. Je to pravdepodobne najbližší možný zážitok k braniu drog bez brania drog – aj keď to neviem porovnať.

To nemôže byť zdravé.

Nie že by som namietal proti eufórii. Znepokojuje ma tá exkluzivita. Prečo by mal byť objav hoden *menej* len preto, lebo niekto *iný* už pozná odpoveď?

Najláskavejšie psychologické vysvetlenie je, že nebudete s jedným problémom zápasiť celé mesiace alebo roky, ak je to niečo, čo si môžete vyhľadať v knižnici. A že ten ohromne povznášajúci pocit vychádza z toho, že ste sa pustili do daného problému z každého možného uhla a narazili ste; potom ste analyzovali problém znovu, použili ste každú myšlienku, čo vám napadla, a všetky údaje, ku ktorým ste sa dostali – postupovali ste po maličkých kúskoch – takže keď ste sa *nakoniec* prelomili cez problém, všetky tie visiace kúsky a nevyriešené otázky zapadli na svoje miesto naraz, ako keby ste vyriešili tucet záhadných vrážd v zamknutých miestnostiach pomocou jediného náznaku.

Navyše, pochopenie, ktoré takto získate, je *skutočným* pochopením – pochopenie, ktoré zahŕňa všetky náznaky, ktoré ste študovali pri riešení problému, keď ste ešte nepoznali odpoveď. Pochopenie, ktoré prichádza z kladenia si otázok deň za dňom a rozmýšľania nad nimi; pochopenie, ktoré nikto iný nebude mať (bez ohľadu na to, koľko im budete hovoriť túto odpoveď) bez toho, aby strávil mesiace štúdiom daného problému v jeho historickom kontexte, aj keď už je vyriešený – a ešte ani vtedy nebude mať zážitok z vyriešenia toho všetkého naraz.

To je jeden možný dôvod, prečo sa James Clerk Maxwell asi viac zabával pri *objavení* Maxwellových rovníc, než vy pri čítaní o nich.

O trochu menej láskavé vysvetlenie je, že tento ohromne povznášajúci pocit pochádza z toho, čo sa v sociálnej psychológii *zdvorilo* nazýva „záväzok“ a „konzistencia“ a „kognitívna disonancia“; tá časť, kde si niečo ceníme viac *iba* preto, lebo nás to stálo viac práce. Štúdie ukazujú, že k čím drsnejšiemu vstupnému rituálu do spolku sa pokusné osoby musia zaviazať, tým viac sú presvedčení o hodnote tohto spolku – rovnaké víno vo fľaši s vyššou uvedenou cenou sa hodnotí ako lepšie chutiace – tento druh vecí.

Samozrejme, ak vás riešenie hlavolamov skrátka baví viac než počutie riešenia, pretože sa vám páči samotná poznávacia práca, na tom nie je nič zlé. Menej láskavým vysvetlením by bolo, keby poplatok 100 dolárov za prezradenie riešenia hlavolamu spôsobil, že by ste považovali toto riešenie za zaujímavejšie, hodnotnejšie, dôležitejšie, prekvapivajšie, atď. než keby ste túto odpoveď dostali zadarmo.

(Mám silné podozrenie, že veľká časť problému PR vedy u širokého obyvateľstva je, že ľudia inštinktívne veria, že keď poznanie rozdávate zadarmo, nemôže byť dôležité. Keby ste museli podstúpiť desivý vstupný rituál, aby vám povedali pravdu o evolúcii, možno by ľudia boli spokojnejší s odpoveďou.)

To naozaj neláskavé vysvetlenie je, že radosť z objavu sa týka spoločenského postavenia. Súťaž. Vzácne statky. Poraziť všetkých ostatných. Nezáleží na tom, či máte 3-izbový alebo 4-izbový dom,

podstatné je mať väčší dom než susedia. Aj 2-izbový dom by bol dobrý, keby ste mali istotu, že susedia budú mať ešte menej.

Nemám principiálne námietky voči súťaži. Nemyslím si, že hra Go je barbarská a mala by byť zakázaná, aj keď je to hra s nulovým súčtom. Ale keby sa eufória z vedeckého objavu *musela* týkať vzácnosti, znamenalo by to, že je dostupná iba jednému človeku z celej civilizácie pre každú konkrétnu pravdu.

Ak je radosť z vedeckého objavovania jednorazová za objav, potom z pohľadu teoretika zábavy Newton pravdepodobne spotreboval výrazné zvýšenie celkovej dostupnej fyzikálnej zábavy za celú históriu pozemského života. Ten sebecký bastard vysvetlil obežné dráhy planét *aj* prílivové vlny.

A situácia je v skutočnosti ešte horšia, pretože podľa štandardného modelu fyziky (objavenom bastardmi, ktorí pokazili hlavolam pre všetkých ostatných) je vesmír priestorovo nekonečný, inflačne sa rozvetvujúci, a rozvetvujúci sa pomocou dekoherencie, čo sú aspoň tri rôzne spôsoby, ako je Skutočnosť exponenciálne alebo nekonečne veľká.

Takže mimozemšťania alebo Newtonovia z alternatívnych svetov alebo iba tegmarkovské duplikáty Newtona mohli všetci objaviť gravitáciu pred *naším* Newtonom – ak veríte, že slovo „pred“ niečo znamená pri takýchto druhoch oddelenia.

Keď mi toto prvýkrát napadlo, v skutočnosti som zistil, že ma to povzbudilo. Keď som si uvedomil, že niekto niekde v šíravách priestoru a času už pozná odpoveď na ľubovoľnú zodpovedateľnú otázku – dokonca aj na otázky z biolóogie a histórie; existujú iné dekoherentné Zeme – uvedomil som si, že aké pochabé bolo myslieť si, že by radosť z objavu mala byť obmedzená na jednu osobu. Stala by sa tak celkom neodvratným zdrojom nevyriešiteľnej existenciálnej úzkosti, čo ja považujem za nezmysel.

Konzistentné riešenie, ktoré zachováva *možnosť* zábavy, je prestať sa znepokojovať ohľadom toho, čo druhí ľudia vedia. Ak nepoznáte odpoveď, je to pre vás tajomstvo. Ak môžete zdvihnúť ruku a zaťať prsty v päsť a netušíte, ako to váš mozog robí – alebo dokonca aké presne svaly máte pod kožou – musíte sa považovať za rovnako nevedomého ako bol lovec-zberač. Iste, niekto iný tú odpoveď pozná – ale aj v časoch lovcov a zberačov niekto iný na alternatívnej Zemi alebo hoci len niekto iný v budúcnosti vedel, aká je odpoveď. Tajomstvo a radosť z nachádzania je buď osobná vec alebo neexistuje vôbec – a ja radšej hovorím, že je to osobné.

Radosť z pomáhania svojej civilizácii povedaním jej niečoho, čo dovtedy nevedela, má sklon byť jednorazová za objav za civilizáciu; tento statok je vzácny, rovnako ako Nobelove ceny. A predstava takejto odmeny môže byť to, čo vás udrží sústredených na jeden problém celé tie roky potrebné na vyvinutie naozaj *hlbokého* porozumenia; plus, keď pracujete na probléme, ktorý vaša civilizácia nepozná, zaručene sa vyhnete predčasnému prečítaniu riešenia.

Ale ako súčasť môjho všeobecného projektu na zrušenie predstavy, že racionalisti majú menej zábavy, chcem obnoviť mágiu a tajomstvo každej časti sveta, ktorej vy *osobne* nerozumiete, bez ohľadu na to, aké iné poznanie môže existovať, ďaleko v priestore a čase, alebo možno aj v mysli vášho suseda. Ak to *vy* neviete, je to tajomstvo. A teraz pomyslite na to, koľko veľa vecí neviete! (Ak si nedokážete na nič spomenúť, máte iné problémy.) Nie je zrazu svet omnoho tajomnejším a magickejším a *zaujímavejším* miestom? Ako keby ste boli premiestnení do alternatívneho vesmíru a museli by ste sa naučiť všetky pravidlá od začiatku?

> „Kamarát mi raz povedal, že sa pozerám na svet, akoby som ho nikdy predtým nevidel. Pomyslel som si, to je pekný kompliment... Moment! Ja som ho *naozaj* predtým nevidel! A čo – dostali všetci ostatní predpremiéru?“

--Ran Prieur

* →

→ http://lesswrong.com/lw/os/joy_in_discovery/

204. Pripútajte sa k skutočnosti

Možno si teda čítate toto všetko a kladiete si otázku: „Áno, ale čo má toto spoločné s redukcionizmom?“

Čiastočne je to otázka ponechania ústupovej línie. Nie je ľahké rozobrať niečo *dôležité* na súčiastky, ak ste presvedčení, že to odoberá mágiu zo sveta, že to pára dúhu. Mám v úmysle rozobrať určité veci na tomto blogu, a radšej by som nevytváral zbytočnú existenciálnu úzkosť.

Čiastočne je to krížová výprava proti Hollywoodskej Rozumnosti, predstave, že porozumenie dúhe odoberá z jej krásy. Dúha je stále krásna, a *navyše* dostanete krásu fyziky.

Ale ešte hlbšie, je to jedna z tých jemných vecí skrytých v jadre rozumnosti. Viete, tie veci, kde začnem hovoriť o „Ceste“. Je to o *pripútaní sa k skutočnosti*.

Ak si dobre pamätám, v jednej z Frankových Herbertových kníh o *Dune* sa píše, že Hovorca Pravdy získava svoju schopnosť odhaliť lži druhých tým, že sám vždy hovorí pravdu, takže si s pravdou vytvorí vzťah, ktoré narušenie dokáže cítiť. Nefungovalo by to, ale aj tak si myslím, že je to jedna z tých krajších predstáv v literatúre. Prinajmenšom, aby ste sa k pravde dostali blízko, musíte byť ochotný pritlačiť sa k skutočnosti tak tesne, ako sa dá, bez cúvnutia alebo úškrnu.

Túto tému pripútania sa k skutočnosti môžete vidieť v kapitole „Lotérie: Plytvanie nádejou“. Porozumenie, že žreb v lotérii má negatívny očakávaný úžitok, neznamená, že prestanete dúfať o tom, že zbohatnete. Znamená to, že prestanete plytvať touto nádejou na žreby v lotérii. Vložíte túto nádej do svoje práce, do svojej školy, do svojej firmy, do svojho vedľajšieho príjmu na eBay; a ak naozaj nemáte nič, čo by vám dávalo nádej, potom je asi načase začať sa obzerať.

Ja nenamietam voči snom, iba voči *nemožným* snom. Lotéria nie je nemožná, ale je takmer nemožne nevykonateľná. Nie preto, že by vyhranie v lotérii bolo extrémne *ťažké* – že by si vyžadovalo zúfalé úsilie – ale preto, že o *prácu* tam nejde.

Toto všetko hovorím, aby som dal príklad myšlienky, že môžeme vziať emocionálnu energiu, ktorá odteká do prázdnoty, a pripútať do ríše skutočnosti.

To neznamená klásť si ciele dosť nízko na to, aby boli „realistické“, čiže ľahké a bezpečné a schválené rodičmi. Možno vo vašom osobnom prípade je to dobrá rada, ja neviem, ale nie je to to, čo hovorím.

Myslím tým, že môžete investovať svoju emocionálnu energiu do dúhy, aj keď sa ukáže, že *nie je* magická. Budúcnosť je vždy absurdná, ale nikdy nie je *neskutočná*.

Hollywoodska Rozumnosť má stereotyp že „rozumný = bez emócií“; že čím rozumnejší ste, tým viac z vašich emócií rozum nevyhnutne zničí. V kapitole „Rozumné cítenie“ som to dal do kontrastu s *„Čo môže byť zničené pravdou, nech je zničené“* a *„Čo pravda podporuje, nech prekvitá“*. Keď ste prišli k svojmu najlepšiemu obrazu pravdy, nie je nič nerozumné na emóciách, ktoré cítite v dôsledku toho – tieto emócie nemôžu byť zničené pravdou, preto nemôžu byť nerozumné.

Takže namiesto *ničenia* emocionálnej energie spojenej so zlými vysvetleniami dúhy, ako by od nás žiadal stereotyp Hollywoodskej Rozumnosti, *presmerujme* túto emocionálnu energiu do skutočnosti – pripútajme ju k názorom, ktoré sú také pravdivé, aké len dokážeme mať.

Chcete lietať? Nevzdávajte sa lietania. Vzdajte sa nápojov na lietanie, a postavte si lietadlo.

Pamätáte na tému „Mysli ako skutočnosť“, kde som hovoril o tom, že keď vám fyzika pripadá antiintuitívna, musíte prijať, že to nie *fyzika* je čudná, ale *vy*?

O čom teraz hovorím, je podobné, ale s emóciami namiesto hypotéz – pripútajte svoje emócie k skutočnému svetu. Nie ku každodennému „realistickému“ svetu. Bol by som vrieskajúci pokrytec, keby som vám povedal, aby ste sklapli a robili si domáce úlohy. Myslím tým *ozajstný* skutočný svet, vesmír, ktorý sa riadi zákonmi, ktorý zahŕňa aj absurdity ako pristátie na Mesiaci a evolúcia ľudskej inteligencie. Akurát tam nie je mágia, nikde, nikdy.

Je mémom Hollywoodskej Rozumnosti, že „Veda uberá životu zábavu.“

Vedá vracia zábavu *do* života.

Rozumnosť smeruje vaše emocionálne energie do vesmíru, namiesto niekam inam.

*→

205. Ak chcete mágiu, mágia vám nepomôže

> Väčšina bosoriek neverila v bohov. Vedeli, že bohovia existujú, samozrejme. Občas s nimi mali do činenia. Ale neverili v nich. Poznali ich príliš dobre. Bolo by to ako veriť v poštára.
>
> --Terry Pratchett, Bosorky na cestách[194]

Jedného dňa som hútal o filozofii fantasy príbehov...

A skôr než ma niekto začne karhať za „neschopnosť pochopiť, o čom je fantasy“, dovoľte mi povedať: bol som vychovaný v SF&F domácnosti. Čítal som fantasy príbehy od veku päť rokov. Občas sa pokúšam *písať* fantasy príbehy. A *nie som* ten typ človeka, ktorý sa pokúša písať v nejakom žánri bez rozmýšľania o jeho filozofii. Odkiaľ si myslíte, že pochádzajú nápady na príbehy?

Každopádne:

Hútal som o filozofii fantasy príbehov, a napadlo mi, že keby v našom svete naozaj existovali draci – keby ste mohli zájsť do zoo, alebo dokonca na vzdialenú horu, a stretli by ste draka chrliaceho oheň – zatiaľ čo nikto by nikdy naozaj nevidel zebru, potom by naše fantasy príbehy obsahovali množstvo zebier, kým draci by nás nevzrušovali.

Tomu hovorím zatlačiť sám seba do kúta, čo? Tráva je vždy zelenšia na opačnej strane neskutočnosti.

V jednej zo štandardných fantasy záplietok sa hlavný hrdina z našej Zeme, sympatická postava s mizernými známkami alebo ťaživou hypotékou, ale stále s dobrým srdcom, zrazu nájde vo svete, kde namiesto vedy funguje mágia. Hlavný hrdina časti začne praktizovať mágiu, a postupom času sa stane (super mocným) čarodejníkom.

Je to však otázka – a áno, je trochu nepríjemná, ale myslím si, že si ju treba položiť: Predpokladám, že väčšina čitateľov týchto románov sa vidí v koži hlavného hrdinu, fantazírujú o svojom vlastnom nadobudnutí čarodejníctva. Túžia po mágii. A pokiaľ to nie je nepravdepodobná demografická skupina, väčšina čitateľov týchto románov nie sú vedci.

Narodili sa vo svete vedy, a nestali sa vedcami. Prečo si myslia, že vo svete mágie by jednali inak?

Ak nemajú vedecký postoj, že nič nie je „iba“ - schopnosť zaujímať sa o iba skutočné veci – ako im pomôže mágia? Keby naozaj *mali* mágiu, bola by iba *skutočná*, a stratila by kúzlo nedosiahnuteľnosti. Mohli by ňou byť spočiatku nadšení, ale (ako víťazi lotérie, ktorí o šesť mesiacov neskôr nie sú takí šťastní, ako očakávali, že budú) toto vzrušenie by časom opadlo. Pravdepodobne hneď ako by museli naozaj *študovať* kúzla.

Pokiaľ by nedokázali nájsť schopnosť tešiť sa z vecí, ktoré sú iba skutočné. Byť rovnako vzrušený zo závesného lietania ako z drakov; byť rovnako vzrušený z vyvorenia svetla pomocou elektriny ako z vytvorenia svetla pomocou mágie... aj keď si to vyžaduje trochu štúdia...

Nechápte ma zle. Ja nekritizujem drakov. Kto vie, možno si jedného dňa nejakých vytvoríme.

Ale ak nemáte schopnosť tešiť sa zo závesného lietania aj keď je *iba skutočné*, potom akonáhle sa draci *stanú* skutočnými, nebudete sa z drakov tešiť o nič viac než zo závesného lietania.

Myslíte si, že by ste radšej žili v Budúcnosti než v prítomnosti? To je celkom pochopiteľná preferencia. Zdá sa, že veci sa časom zlepšujú.

Ale nezabudnite, že *toto je* Budúcnosť v porovnaní so stredovekom pred tisíc rokmi. Máte príležitosti, o ktorých vtedy nesnívali ani králi.

Ak tento trend bude pokračovať, Budúcnosť veru môže byť veľmi dobrým miestom na život. Ale ak sa dostanete do Budúcnosti, až tam budete, nájde tam iba ďalšie Dnes. Ak nemáte tú základnú schopnosť

→ http://lesswrong.com/lw/ot/bind_yourself_to_reality/

194 Terry Pratchett, *Witches Abroad* (London: Corgi Books, 1992).

užívať si Dnes – ak vaša emocionálna energie dokáže ísť *iba* do Budúcnosti, ak dokážete *iba* dúfať v lepší zajtrajšok – potom vám nedokáže pomôcť žiadne plynutie času.

(Áno, v Budúcnosti by mohla existovať tabletka, ktorá napraví emocionálny problém hľadenia stále do Budúcnosti. Nemyslím si, že to vyvracia moju pointu, ktorá je to tom, aký druh tabletiek by sme mali chcieť brať.)

Matthew C., diskutujúci tu na *Less Wrong*, sa zdá byť veľmi nadšený neformálne špecifikovanou „teóriou" Ruperta Sheldrakea, ktorá „vysvetľuje" také javy nevyžadujúce si vysvetľovanie, ako je skladanie proteínov a súmernosť snehových vločiek. Ale prečo nie je Matthew C. rovnako nadšený povedzme špeciálnou relativitou? O špeciálnej relativite naozaj *vieme*, že je to zákon, prečo teda nie je *ešte* vzrušujúcejšia? Nadšenie zo zákona, o ktorom sa už vie, že je pravdivý, má tú výhodu, že viete, že vaše nadšenie nevyjde nazmar.

Keby bola Sheldrakeova teória akceptovanou pravdou, ktorá sa učí na základných školách, Matthew C. by sa o ňu nestaral. Prečo inak by bol Matthew C. fascinovaný týmto jedným konkrétnym zákonom, o ktorom verí, že je fyzikálny zákon, viac než všetkými ostatnými zákonmi?

Najhoršia pohroma, ktorá by mohla postihnúť komunitu New Age, by bola, keby ich rituály začali spoľahlivo fungovať, a keby sa UFO naozaj objavovali na oblohe. Aký zmysel by malo veriť v mimozemšťanov, keby tam skrátka *boli*, a keby ich videli aj všetci ostatní? Vo svete, v ktorom by psychotronické sily boli iba skutočné, priaznivci New Age by *neverili* v psychotronické sily, rovnako ako nikomu nezáleží na gravitácii natoľko, aby v ňu veril. (Okrem vedcov, samozrejme.)

Prečo som taký negatívny vočí mágii? Bolo by *zlé*, keby mágia existovala?

V skutočnosti nie som negatívny vočí mágii. Pamätajte, že občas píšem fantasy príbehy. Ale otravuje ma psychológia, ktorá keby sa narodila vo svete, kde kúzla a nápoje fungujú, by túžila po svete, v ktorom sa predmety domácej spotreby vyrábajú v hojnosti na bežiacich pásoch.

Časť pripútania sa k skutočnosti, na emocionálnej rovnako ako na intelektuálnej úrovni, je zmierenie sa s faktom, že *žijete tu*. Iba potom dokážete vidieť toto, svoj svet, a všetky príležitosti, ktoré vám poskytuje, bez želania, aby ste to nevideli.

Aby som to nerobil príliš abstraktným, ja vo svojom rodnom svete *nemám* nedostatok drakov na bojovanie, ani mágie na ovládnutie. Keby som bol ja premiestnený do jedného z týchto fantasy románov, neprekvapilo by mal, keby som zistil, že študujem zakázané najvyššie kúzla...

...pretože prečo by malo premiestnenie do magického sveta niečo zmeniť? Nejde o to, *kde* ste, ale *kto* ste.

Pamätajte si teda Litániu proti premiestneniu do alternatívneho vesmíru:

Ak mám byť niekde šťastný,
Alebo niekde dosiahnuť slávu,
Alebo sa niekde dozvedieť skutočné tajomstvá,
Alebo niekde zachrániť celý svet,
Alebo niekde prežívať silné pocity,
Alebo niekde pomáhať ľuďom,
Môžem to rovnako dobre robiť v skutočnosti.

*→

206. Všedná mágia

Ako si možno spomínate spred pár mesiacov, myslím si, že časťou racionalistovho étosu je *pripútať sa emocionálne* k absolútne zákonitému redukcionistickému vesmíru – vesmíru, ktorý neobsahuje žiadne ontologicky základné myšlienkové veci ako sú duše alebo mágia – a vliať všetku svoju nádej a všetku svoju starosť do tohto iba skutočného vesmíru a jeho možností, bez sklamania.

→ http://lesswrong.com/lw/ou/if_you_demand_magic_magic_wont_help/

Existuje starý trik ako bojovať s nespokojnosťou, kde si urobíte zoznam vecí, za ktoré ste vďační, ako je napríklad strecha nad vašou hlavou.

Prečo si teda neurobiť zoznam schopností, ktoré by boli úžasne cool, *keby boli magické*, alebo keby ich malo iba pár vybraných jednotlivcov?

Predstavte si napríklad, že by ste namiesto jedného oka vlastnili magické *druhé* oko umiestnené na čele. A toto druhé oko by vám umožňovalo *vidieť v treťom rozmere* – takže by ste dokázali nejako povedať, *ako ďaleko* sú rôzne veci – kým obyčajné oko by videlo iba dvojrozmerný tieň skutočného sveta. Iba vlastníci tejto schopnosti dokážu presne zamieriť legendárne zbrane zabíjajúce na vzdialenosť väčšiu než meč, alebo využiť plný potenciál škrupín ultrarýchlych mechanizmov nazývaných „autá".

„Binokulárne videnie" by bolo pre takúto schopnosť príliš slabým pojmom. Ocenili by sme ju až keby mala príslušne pôsobivé meno, napríklad Mystické Oči Vnímania HĺbkyTu je teda zoznam niektorých z mojich obľúbených magických schopností:

- *Vibračná Telepatia.* Prenášaním neviditeľných vĺn cez samotný vzduch si môžu dvaja vlastníci tejto schopnosti *vymieňať myšlienky*. V dôsledku toho môže Vibračná Telepatia vytvárať emocionálne putá omnoho hlbšie než je možné u iných primátov.
- *Psychometrické Stopy.* Zaznamenávaním malých tenkých stôp na povrchu dokáže Psychometrický Stopár zanechať odtlačok emócií, dejín, vedomostí, dokonca aj štruktúru iných kúziel. Toto je vyššia úroveň než Vibračná Telepatia, pretože Psychometrický Stopár dokáže zdieľať myšlienky dávno mŕtvych Stopárov, ktorí žili pred tisíckami rokov. Čítaním jednej Stopy a zároveň zakresľovaním druhej dokážu Stopári zdvojovať Stopy; a tieto zdvojené Stopy môžu dokonca obsahovať podrobné vzory iných kúziel a mágie. Takto Stopári disponujú takmer nepredstaviteľnou magickou silou; môžu sa však dostať do problémov pri snahe používať komplikované Stopy, ktoré by nedokázali sami vystopovať.
- *Multidimenzionálna Kinéza.* Pomocou jednoduchých, takmer bezmyšlienkovitých aktov vôle môžu Kinetici spôsobiť, že mimoriadne zložité sily prejdú cez ich malé chápadlá do ľubovoľného hmotného predmetu v ich dosahu – nie iba tlačenie, ale aj kombinácia tlačenia v mnohých bodoch, ktorá môže účinne aplikovať stláčanie a krútenie. Táto Kinetická schopnosť je omnoho jemnejšia než sa na prvý pohľad zdá: dokážu ju používať nielen na ovládanie existujúcich predmetov so smrtiacou presnosťou, ale aj na používanie síl, ktoré pretvárajú predmety do tvarov ešte vhodnejších na Kinetické ovládanie. Dokážu dokonca vytvoriť nástroje, ktoré rozširujú moc ich Kinézy a umožňujú im pretvárať čoraz jemnejšie a čoraz zložitejšie nástroje, čo je pozitívna spätná väzba rovnako pôsobivá ako znie.
- *Oko.* Používateľ tejto schopnosti dokáže vnímať nekonečne malé putujúce uzlíky Sily, ktorá spája hmotu – drobné vibrácie podobné životodarnej sile Slnka dopadajúcej na listy, ale omnoho jemnejšie. Majiteľ Oka dokáže vnímať predmety ďaleko za hranicou svojho dotyku, pomocou drobných porúch, ktoré spôsobujú v tejto Sile. Hory vzdialené mnoho dní putovania sú im známe akoby boli na dosah ruky. Podľa majiteľov Oka, keď prichádza noc a slabne slnečné svetlo, vnímajú obrovský oheň planúci v nepredstaviteľnej vzdialenosti – hoci nikto iný nemá žiaden spôsob, ako to overiť. Hovorí sa, že vlastníctvo jedného Oka robí jeho majiteľa ekvivalentom vládcu.

A nakoniec,

- *Najvyššia Moc.* Majiteľ tejto schopnosti obsahuje menšiu, nedokonalú ozvenu celého vesmíru, čo mu umožňuje vyhľadávať cesty v pravdepodobnosti k ľubovoľnej želanej budúcnosti. Ak vám to znie ako absurdne mocná schopnosť, máte pravdu – herná rovnováha po tomto letí priamo z okna. Je mimoriadne zriedkavá medzi živými formami, je to *sekai no ougi*, čiže „skrytá technika sveta".

Nič sa nemôže vzoprieť Najvyššej Moci, okrem Najvyššej Moci. Ľubovoľná menšia než najvyššia moc bude tou Najvyššou jednoducho „pochopená" a prerušená nejakým nepredstaviteľným spôsobom, alebo dokonca pohltená do repertoáru Najvyššej. Z tohto dôvodu na Najvyššia Moc niekedy nazýva „majstrovská technika všetkých techník" alebo „tromfová karta, ktorá tromfne

všetky zvyšné tromfy“. Tí mocnejší Najvyšší dokážu svoje „pochopenie“ rozprestieť cez galaktické vzdialenosti a eóny času, a dokonca vnímať čudesné zákony skrytého „sveta pod týmto svetom“.

Najvyšší už boli zabití obrovskými prírodnými katastrofami alebo extrémne rýchlymi prekvapivými útokmi, ktoré im nedali šancu použiť svoju moc. Ale všetky takéto víťazstvá sú v konečnom dôsledku otázkou šťastia – nedokážu vzdorovať Najvyšším na ich vlastnej úrovni ohýbania pravdepodobnosti, a ak prežijú, začnú ohýbať Čas, aby sa vyhli budúcim útokom.

Lenže samotná Najvyššia Moc je tiež nebezpečná, a mnohí Najvyšší už boli zničení svojou vlastnou mocou – upadli do jednej z chýb v ich nedokonalej vnútornej ozvene sveta.

Hoci by bol zbavený zbraní a brnenia, a zamknutý v kobke, Najvyšší je stále jednou z najnebezpečnejších foriem života na celej planéte. Meč možno zlomiť, a končatinu možno odťať, ale Najvyššia Moc je „moc, ktorú nemožno zničiť, dokiaľ nezničia teba“.

Azda preto, že toto spojenie je také dôverné, Najvyšší považujú tých, ktorí trvalo stratia svoju Najvyššiu Moc – bez nádeje na jej znovuzískanie – za schiavo, čiže „dýchajúcich mŕtvych“. Najvyšší tvrdia, že Najvyššia Moc je taká dôležitá, že je nevyhnutnou časťou toho, čo robí tvora cieľom namiesto prostriedkom. Najvyšší dokonca trvajú na tom, že ten, kto nemá Najvyššiu Moc, nedokáže ani len začať naozaj rozumieť Najvyššej Moci, a teda nedokáže pochopiť, prečo je Najvyššia Moc taká morálne dôležitá – čo je podozrivo výhodný argument.

Majitelia tejto schopnosti tvoria absolútnu aristokraciu a zaobchádzajú so všetkými ostatnými formami života ako so svojimi hračkami.

*→

207. Krása ustálenej vedy

Fakty nemusia byť nevysvetliteľné, aby boli krásne; pravdy sa nestanú menej hodnými poznania, ak ich vie niekto iný; názory sa nestanú menej hodnotnými, ak ich majú aj mnohí iní...

...a ak sa zaujímate iba o vedecké témy, ktoré sú kontroverzné, skončíte s hlavou plnou odpadu.

Médiá si myslia, že iba najnovšie vedecké výsledky si zaslúžia reportáže. Ako často vidíte titulky ako: „Všeobecná relativita stále riadi obežné dráhy planét“ alebo „Teória flogistonu zostáva nepravdivou“? Teda keď už je niečo solídna veda, už to nie je prevratná titulka. Veda, ktorá „stojí za správu v novinách“ je často založená na minime indícií a v polovici prípadov nesprávna – keby to nebolo na najvzdialenejších končinách vedy, neboli by to prevratné novinky.

Vedecké *kontroverzie* sú problémom takým *náročným*, že ešte aj ľudia, ktorí strávili roky zvládaním danej oblasti, sa stále dokážu sami napáliť. To vytvára tie zapálené hádky, ktoré priťahujú pozornosť médií.

Čo je horšie, ak nie ste v danej oblasti a nezúčastňujete sa hry, kontroverzie *nie sú ani zábavné*.

Ach, iste, môžete sa zabaviť tým, že si vyberiete nejakú stranu v spore. Ale to môžete získať aj v hocijakom futbalovom zápase. To nie je ten druh zábavy, o ktorom je veda.

Keď si prečítate dobre napísanú učebnicu, dostanete: Starostlivo sforumované vysvetlenia pre začínajúcich študentov, matematiku odvodenú krok za krokom (kde je to vhodné), množstvo citovaných ilustračných pokusov (kde je to vhodné), cvičné problémy na predvedenie svojej novej zručnosti, a pomerne dobrú záruku, že čo sa učíte, je naozaj pravda.

Pri čítaní novinových správ zvyčajne dostanete: Falošné vysvetlenia, ktoré vám nedajú nič okrem ilúzie pochopenia výsledku, ktorému autor novinovej správy sám nerozumel, a ktorý má pravdepodobne šancu pol na pol, že sa ho nepodarí replikovať.

Moderná veda je založená na objavoch, ktoré sú založené na objavoch, ktoré sú založené na objavoch, a tak ďalej, až späť k ľuďom ako bol Archimedes, ktorý objavil fakty ako prečo lode

→ http://lesswrong.com/lw/ve/mundane_magic/

plávajú, ktoré dávajú zmysel aj keď neviete o iných objavoch. Dobré miesto na začiatok putovanie po tejto trase je na jej začiatku.

Nehanbite sa ani prečítať si vedecké učebnice *pre začiatočníkov*. Ak chcete predstierať, že ste sofistikovaný, nájdite si nejaký zápas, na ktorom môžete pískať. Ak sa skrátka chcete zabávať, pamätajte, že jadrom vedeckej krásy je jednoduchosť.

A myslieť si, že môžete skočiť priamo na hranice, keď ste sa ešte nenaučili ustálenú vedu, je ako...

...ako skúšať vyliezť iba *hornú* polovicu Mount Everestu (lebo to je jediná časť, ktorá vás zaujíma) tak, že sa postavíte pod horu, zohnete kolená, a *naozaj silno* vyskočíte (aby ste preskočili tie nudné časti).

Tým nehovorím, že by ste nemali venovať pozornosť vedeckým kontroverziám. Ak si 40 % onkológov myslí, že biele ponožky spôsobujú rakovinu, a zvyšných 60 % zúrivo nesúhlasí, je dôležité o tom vedieť.

Akurát si nemyslite, že veda *musí* byť kontroverzná, aby bola zaujímavá.

Alebo, keď už sme pri tom, že veda musí byť nedávna, aby bola zaujímavá. Stála diéta vedeckých *noviniek* vám škodí: Ste to, čo jete, a ak konzumujete iba vedecké správy o pohyblivých oblastiach, bez občasnej pevnej učebnice, váš mozog sa premení na rôsol.

* →

208. Deň úžasných objavov: Prvý apríl

Myslíte si: „Prvý apríl... nemá to náhodou byť deň bláznov?“

Áno – a to poskytne ideálnu zámienku na oslavovanie Dňa úžasných objavov.

Ako som hovoril v kapitole „Krása ustálenej vedy“, je veľký problém, že médiá sa z oblasti vedy sústredia iba na *prevratné novinky*. Prevratné novinky sa vo vede odohrávajú na najvzdialenejších kútoch vedeckej hranice, čo znamená, že nový objav je často:

- Kontroverzný
- Podporený iba jediným experimentom
- Omnoho komplikovanejší než bežný smrteľník dokáže zvládnuť, a vyžaduje si veľa predchádzajúcej vedy na pochopenie, čo je dovod, prečo to nebolo vyriešené už pred tristo rokmi
- Neskôr sa ukáže ako nesprávny

Ľudia nikdy neuvidia tie *solídne* veci, a už vôbec nie tie *zrozumiteľné* veci, pretože to nie je *prevratná novinka*.

Na Deň úžasných objavov však navrhujem, aby novinári, ktorým na vede naozaj záleží, písali – pod ochranným štítom prvého apríla – o takých dôležitých, ale ignorovaných vedeckých príbehoch ako:

- VYSVETLENÉ LODE: Nudista vo vani vyriešil stáročia starý problém
- NEPREJDEŠ! Zmarené nádeje turistov v Kaliningrade
- MÁTE V PĽÚCACH *OHEŇ*? Vedci postupne pripúšťajú súvislosť medzi dýchaním a spaľovaním

Všimnite si, že každá z týchto tituliek je *pravdivá* – opisuje udalosti, ktoré sa naozaj stali. Akurát sa nestali *včera*.

V histórii vedy bolo veľa ľudsky pochopiteľných úžasných prevratných objavov, ktoré dokážete pochopiť aj bez PhD alebo dokonca bez bakalára. Kľúčovým slovom je tu *história*. Pomyslite na Archimedovo „Heuréka!“, keď pochopil vzťah medzi vodou, ktorú loď vytlačí, a dôvodom, prečo loď pláva. To je *dosť ďaleko vzadu* v dejinách vedy, takže nepotrebujete poznať ďalších 50 objavov, aby ste pochopili teóriu; dá sa vysvetliť pár grafmi; každý vidí, na čo je to užitočné; a potvrdzujúci experiment dokážete duplikovať vo svojej vlastnej vani.

Moderná veda je založená na objavoch založených na objavoch založených na objavoch a tak ďalej až späť k Archimedovi. Písať vedecké reportáže *iba* o prevratných novinách je ako prísť do kina v 3/4 filmu, napísať príbeh o tom, ako „Muž s krvavými rukami pobozká dievča držiace pištoľ!“, a opäť odísť.

→ http://lesswrong.com/lw/ow/the_beauty_of_settled_science/

A keby váš redaktor povedal: „Ach, ale to našich čitateľov nebude zaujímať...“

Potom mu ukážte, že Reddit a Digg neodkazujú *iba* na prevratné novinky. Odkazujú aj na krátke webové stránky, ktoré dávajú dobré vysvetlenia starej vedy. Čitatelia za ne hlasujú, a to by vám malo niečo povedať. Vysvetlite, že ak sa vaše noviny nechcú zmeniť, aby vyzerali viac ako Reddit, budete musieť začať predávať drogy, aby ste mali na výplaty. Redaktori takéto veci radi počúvajú, však?

Na internete dobré nové vysvetlenie starej vedy *je* novinkou a šíri sa ako novinka. Prečo by nemohli vedecké časti novín fungovať rovnako? Prečo nie je nové *vysvetlenie* hodné reportáže?

Ale toto všetko je na prvý krok príliš vizionárske. Zatiaľ si počkajme, či sa niektorí novinári tam vonku podujmú na Deň úžasných objavov, kde môžete písať o nejakom *zrozumiteľnom* vedeckom objave, akoby sa to práve stalo.

Prvý apríl. Poznačte si to do kalendára.

*→

209. Je humanizmus náhradou náboženstva?

Mnoho rokov pred bratmi Wrightovcami ľudia snívali o lietaní pomocou čarovných nápojov. Na *čírej túžbe* lietať nebolo nič nerozumné. Na želaní hľadieť zhora na oblaky nebolo nič *skazené*. Akurát tá časť o „čarovných nápojoch“ bola nerozumná.

Predstavte si, že by ste ma dalo do skenera fMRI a urobili film z úrovní aktivity môjho mozgu, kým by som pozeral štart raketoplánu. (Chcieť navštíviť vesmír nie je „realistické“, ale je to v podstate legálny sen – dá sa splniť vo vesmíre riadenom zákonmi.) Možno by fMRI – možno áno, možno nie – pripomínalo fMRI oddaného kresťana sledujúceho betlehem.

Keby experimentátor získal takýto výsledok, je veľa ľudí, kresťanov aj ateistov, ktorí by sa pobavili: „Haha, vesmírne lety sú tvoje náboženstvo!“

Ale to je nesprávne kreslenie hranice kategórií. Je to ako povedať, že keďže sa niektorí ľudia raz pokúšali lietať pomocou nerozumných prostriedkov, nikto by si nikdy nemal užívať pohľad z okna lietadla na oblaky pod ním.

Ak je štart rakety to, čo mi dáva pocit estetickej transcendencie, nevidím to ako *náhradu* za náboženstvo. To je teomorfizácia – názor pobavených religionistov, ktorí predpokladajú, že každý, kto *nie je* veriaci, má v mysli dieru, ktorú potrebuje zaplniť.

Teda, aby som bol férový k religionistom, nejde *iba* o pobavený predpoklad. *Existujú* ateisti, ktorí majú vo svojich mysliach dieru v tvare náboženstva. *Videl som* pokusy nahradiť náboženstvo ateizmom alebo dokonca transhumanizmom. A výsledok je vždy hrozný. Naprosto hrozný. Absolútne biedne hrozný.

Označujem takéto úsilia za „chválospevy na neexistenciu Boha“.

Keď sa niekto rozhodne napísať ateistický chválospev - „Zdravas, ó neinteligentný vesmír“, bla bla bla – výsledok bude, bez výnimky, otrasný.

Prečo? Pretože je to napodobenina. Pretože nemajú žiadnu motiváciu napísať chválospev *okrem* hmlistého pocitu, že keďže kostoly majú chválospevy, aj oni by mali mať nejaký. A to ich z čisto umeleckého hľadiska stavia hlboko pod úroveň skutočného náboženského umenia, ktoré nie je napodobeninou ničoho, ale originálnym vyjadrením emócie.

Náboženské chválospevy (často) napísali ľudia, ktorí to *silno cítili* a *úprimne písali* a venovali veľké úsilie prozódii a obraznosti svojho diela – to je to, čo ich práci dáva jej pôvab a umeleckú integritu.

Sú teda ateisti odsúdený byť bez chválospevov?

Existuje skúšobný kameň pokusov o post-teizmus. Tento skúšobný kameň je: „Keby náboženstvo nebolo nikdy existovalo u ľudského druhu – keby sme *nikdy neboli urobili* tú prvotnú chybu – mala by táto pieseň, táto maľba, tento rituál, tento spôsob myslenia, stále zmysel?“

→ http://lesswrong.com/lw/ox/amazing_breakthrough_day_april_1st/

Keby ľudstvo nebolo nikdy urobilo tú prvotnú chybu, neexistovali by žiadne chválospevy na neexistenciu Boha. Ale stále by existovali manželstvá, takže predstava ateistického sobášneho obradu dáva zmysel – pokiaľ sa náhle nepustíte do lekcie o tom, že Boh neexistuje. Pretože vo svete, kde by náboženstvo *nikdy nebolo existovalo*, by nikto neprerušoval sobášny obrad, aby povedal o neuveriteľnosti vzdialeného hypotetického pojmu. Hovoril by o láske, o deťoch, o záväzku, o úprimnosti, o oddanosti, ale kto do čerta by spomínal Boha?

A v ľudskom svete, kde by náboženstvo *nikdy nebolo existovalo*, by stále boli ľudia, ktorí by mali slzy v očiach pri sledovaní štartu raketoplánu.

Čo je dôvod, prečo aj keby experiment ukázal, že sledovanie štartu raketoplánu rozsvecuje v mojom mozgu oblasti spojené s „náboženstvom“, spojené s pocitmi transcendencie, nevidel by som to ako *náhradu* za náboženstvo; očakával by som, že by sa rozsvietili tie isté oblasti v mozgu, z toho istého dôvodu, keby som žil vo svete, kde náboženstvo nikdy nebolo vynájdené.

Dobrý „ateistický chválospev“ je jednoducho pieseň o niečom, o čom je hodno spievať, čo zhodou okolností nie je náboženské.

Takisto, otočená hlúposť nie je inteligencia. Najväčší blázon na svete môže povedať, že Slnko svieti, ale to nespôsobí, že zhasne. Pointa *nie je* vytvoriť život, ktorý pripomína náboženstvo tak málo ako sa len dá v ľubovoľnom povrchnom aspekte – to je rovnaký spôsob myslenia, aký inšpiruje chválospevy na neexistenciu Boha. Keby ľudstvo nikdy nebolo urobilo tú pôvodnú chybu, nikto by sa *nepokúšal vyhnúť* veciam, ktoré hmlisto pripomínajú náboženstvo. Verte presne, a potom cíťte primerane: Ak vesmírne lety naozaj existujú, a ak pohľad na stúpajúcu raketu vo vás vzbudzuje chuť spievať, potom tú pieseň dočerta zložte.

Ak mám slzy v očiach pri štarte raketoplánu, neznamená to, že sa snažím zaplniť dieru po náboženstve – znamená to, že moja emotionálna energia, môj *záujem*, je pripútaný k skutočnému svetu.

Keby Boh jasne rozprával a spoľahlivo odpovedal na modlitby, stal by sa skrátka ďalšou nudnou obyčajnou vecou, o nič viac hodnou viery než poštár. Keby Boh bol skutočný, zničilo by to tú vnútornú neistotu, ktorá ako kompenzáciu vyvoláva zapálenie. A keby všetci ostatní verili, že Boh je skutočný, zničili by to špeciálnosť bytia jedným z vyvolených.

Ak investujete svoju emocionálnu energiu do kozmických letov, nemáte tieto zraniteľné miesta. Môžem *vidieť* štart Space Shuttle bez straty úžasu. Každý môže veriť, že Space Shuttle je skutočný, a nerobí ho to o nič menej zvláštnym. Ja som sa nezatlačil do kúta.

Voľba medzi Bohom a ľudstvom nie je iba výber drogy. Dôležité je, že ľudstvo *naozaj existuje*.

* →

210. Vzácnosť

Nasledujúce je prebrané najmä z knihy Roberta Cialdiniho *Vplyv: Psychológia presviedčania*.[195] Mám z tejto knihy tri kópie, jednu pre seba, a dve na požičiavanie kamarátom.

Vzácnosť, ako sa tento pojem používa v sociálnej psychológii, je keď sa veci stávajú *viac žiadúce* ako vyzerajú *menej dosiahnuteľné*.

- Ak dáte dvojročného chlapca do miestnosti s dvoma hračkami, jedna voľne položená a druhá za stenou z priesvitného plastu, dvojročný chlapec bude ignorovať tú ľahko dostupnú hračku a pôjde za tou napohľad zakázanou. Ak je stena dosť nízka a dá sa ľahko preliezť, batoľa nedáva prednosť žiadnej z hračiek pred tou druhou.[196]

→ http://lesswrong.com/lw/oy/is_humanism_a_religionsubstitute/

195 Robert B. Cialdini, *Influence: The Psychology of Persuasion: Revised Edition* [Vplyv: Psychológia presviedčania: Upravené vydanie] (New York: Quill, 1993).

196 Sharon S. Brehm and Marsha Weintraub, „Physical Barriers and Psychological Reactance: Two-year-olds' Responses to Threats to Freedom,“ [Fyzické bariéry a psychologická reaktancia: Reakcie dvojročných na ohrozenie slobody] *Journal of Personality and Social Psychology* 35 (1977): 830–836.

- Keď okres Dade zakázal používanie fosfátových pracích prostriedkov, mnohí obyvatelia Dade cestovali do susedných okresov a kúpili obrovské množstvá fosfátových pracích prostriedkov. V porovnaní s obyvateľmi Tampy, ktorých sa regulácia netýkala, obyvatelia Dade hodnotili fosfátové pracie prostriedky ako jemnejšie, účinnejšie, silnejšie voči škvrnám, a dokonca verili že sa ľahšie sypú.[197]

Podobne informácia, ktorá vyzerá zakázaná alebo tajná, vyzerá dôležitejšie a dôveryhodnejšie:

- Keď sa študenti univerzity v Severnej Caroline dozvedeli, že bola zakázaná prednáška proti zmiešaným internátom, začali byť viac proti zmiešaným internátom (hoci tú prednášku nikdy nepočuli).[198]
- Keď vodič povedal, že má poistenie zodpovednosti, experimentálni porotcovia prisúdili jeho obeti v priemere o štyri tisíc dolárov viac než keď vodič povedal, že nemá poistenie. Ak sudca následne informoval porotcov, že informácia o poistení je neprípustná a musí byť ignorovaná, porotcovia prisúdili v priemere o trinásť tisíc dolárov viac než keď vodič povedal, že nemá poistenie.[199]
- Nákupcovia pre supermarkety, ktorým dodávateľ povedal, že je nedostatok hovädzieho mäsa, objednali dvakrát viac hovädzieho mäsa než dodávatelia, ktorým povedal, že je ľahko dostupné. Nákupcovia, ktorým bolo povedané, že je nedostatok hovädzieho mäsa a že samotná táto informácia o nedostatku je ťažko dostupná – že sa o tomto nedostatku všeobecne nevie – objednali šesťkrát viac hovädzieho mäsa. (Keďže štúdia sa konala v kontexte skutočného sveta, poskytnutá informácia bola vlastne pravdivá.)[200]

Konvenčná teória na vysvetlenie tohto javu je „psychologická reaktancia“, čo v jazyku sociálnej psychológie znamená: „Keď povieš ľuďom, že niečo nemôžu robiť, budú sa snažiť o to viac.“ Zdá sa, že základné inštinkty za tým sú zachovanie postavenia a zachovanie možností. Vzdorujeme dominancii, keď sa nejaký ľudský činiteľ snaží obmedziť našu slobodu. A keď hrozí, že prídeme o nejaké možnosti, hoci aj z prirodzených dôvodov, snažíme sa vrhnúť na tieto možnosti skôr než zmiznú.

Vrhanie sa na miznúce možnosti môže byť dobrou adaptáciou v spoločnosti lovcov a zberačov – zbierajte ovocie, kým je ešte na strome – ale v spoločnosti založenej na peniazoch nás môže vyjsť dosť draho. Cialdini hlási, že v jednom obchode s prístrojmi pozoroval predavača, ktorý keď videl, že zákazník prejavuje záujem o nejaký prístroj, prišiel a smutne informoval zákazníka, že tento prístroj je vypredaný, že posledný sa predal pred dvadsiatimi minútami. Vzácnosť vyvolala náhly skok túžby, a zákazník sa často opýtal, či je nejaká možnosť, že by mu predavač našiel nejaký nepredaný kus v sklade alebo niekde. „Dobre,“ povedal predavač, „dá sa to a som ochotný sa pozrieť; ale rozumiem správne, že toto je ten model, ktorý chcete, a keď ho nájdem za túto cenu, vezmete si ho?“

Ako Cialdini poznamenáva, hlavným znakom tohto zlyhania je, že snívate o tom, ako niečo *vlastníte*, namiesto o tom, ako to *používate*. (Timothy Ferriss dáva podobnú radu na plánovanie života; pýtajte sa, aké *opakované zážitky* by vás urobili šťatnými a nie aký majetok alebo zmena postavenia.)

Ale skutočne základný problém s túžbou po nedosiahnuteľnom je, že akonáhle to naozaj *získate*, prestane to byť nedosiahnuteľné. Ak sa nedokážeme tešiť z iba dostupného, naše životy budú *vždy* frustrujúce...

---

197 Michael B. Mazis, Robert B. Settle, and Dennis C. Leslie, „Elimination of Phosphate Detergents and Psychological Reactance,“ [Odstraňovanie fosfátových čistiacich prostriedkov a psychologická reaktancia] *Journal of Marketing Research* 10 (1973): 2; Michael B. Mazis, „Antipollution Measures and Psychological Reactance Theory: A Field Experiment,“ [Opatrenia proti znečisťovaniu a teória psychologickej reaktancie: Prirodzený experiment] *Journal of Personality and Social Psychology* 31 (1975): 654–666.

198 Richard D. Ashmore, Vasantha Ramchandra, and Russell A. Jones, „Censorship as an Attitude Change Induction,“ [Cenzúra ako podnet k zmene postoja] *Článok predložený na stretnutí Eastern Psychological Association* (1971).

199 Dale Broeder, „The University of Chicago Jury Project,“ [Projekt poroty na univerzite v Chicagu] *Nebraska Law Review* 38 (1959): 760–774.

200 A. Knishinsky, „The Effects of Scarcity of Material and Exclusivity of Information on Industrial Buyer Perceived Risk in Provoking a Purchase Decision“ [Účinky vzácnosti materiálu a exkluzívnosti informácie na vnímanie rizika pri vyvolávaní nákupných rozhodnutí u priemyselných nákupcov] (Dizertačná práca, Arizona State University, 1982).

*→

211. Posvätné obyčajno

Tak som si čítal knihu (asi prvú polovicu) *Večný oheň* od Adama Franka,[201] ako som sa pripravoval na svoj rozhovor s ním v Bloggingheads. Kniha Adama Franka je o zážitku posvätného. Ja to tak zvyčajne nenazývam, ale samozrejme viem, o ktorom zážitku Frank hovorí. Je to to, čo ja cítim, keď sledujem video štartu raketoplánu; alebo čo cítím – v menšej miere, pretože v tomto svete je to príliš bežné – keď sa v noci pozerám na hviezdy a myslím na to, čo znamenajú. Alebo na narodenie dieťaťa, povedzme. To, čo je dôležité v Rozvíjaní Príbehu.

Adam Frank tvrdí, že tento zážitok je niečo, čo má veda hlboko spoločné s náboženstvom. Že to nie je napríklad základná ľudská schopnosť, ktorú by náboženstvo kazilo.

Večný oheň cituje *Odrody náboženského zážitku* od Williama Jamesa, ako hovorí:

> Náboženstvo... bude pre nás znamenať pocity, skutky a zážitky jednotlivých ľudí v ich samote; dokiaľ rozumejú, že stoja sami vo vzťahu s tým, čo považujú za nadprirodzené.

A táto téma sa ďalej rozvíja: Posvätnosť je niečo silno *súkromné* a *individuálne*.

Čo na mňa vôbec neurobilo dojem. Nemal by som mať pocity posvätnosti, ak som jeden z *mnohých*, ktorí sledujú video, ako *SpaceShipOne* vyhráva X-Prize? Prečo nie? Mal by som si myslieť, že môj zážitok posvätnosti sa musí nejako *odlišovať* od zážitkov všetkých *ostatných* sledujúcich? Prečo, keď máme všetci rovnaký dizajn mozgu? Veru, načo by som *potreboval* veriť, že som jedinečný? (Ale „jedinečný“ *je* ďalšie slovo, ktoré Adam Frank používa; ten a ten mal „jedinečný zážitok posvätnosti“.) Je tento pocit súkromný v rovnakom zmysle, ako máme problém komunikovať *ľubovoľný* zážitok? Prečo to potom zdôrazňovať pri posvätnosti, a nie pri kýchaní?

Svitlo mi, keď som si uvedomil, že sa pozerám na trik epistemológie Temnej Strany – ak niečo vyhlásite za *súkromné*, ochránili ste to pred kritikou. Môžete povedať: „Nemôžete ma kritizovať, pretože toto je môj súkromný, vnútorný zážitok, ku ktorému sa nikdy nedostatene, aby ste ho mohli spochybniť.“

Ale cenou za ochranu pred kritikou je, že ste vyhnaní do samoty – do samoty, ktorú William James obdivuje ako jadro náboženskej skúsenosti, akoby osamelosť bola *dobrá* vec.

Takéto pozostatky epistemológie Temnej Strany sú kľúčom k porozumeniu mnohých spôsobov, akými náboženstvo prekrúca zážitok posvätnosti:

Tajomnosť – ale prečo by posvätné malo byť tajomné? Štart raketoplánu funguje úplne fajn aj bez tajomnosti. O čo *menej* by som obdivoval hviezdy, keby som *nevedel*, čo sú, iba že sú to malé bodky na nočnej oblohe? Ale ak spochybňujú vaše náboženské názory – keď sa niekto opýta: „Prečo Boh nevylieči mrzákov?“ - potom môžete nájsť útočisko a povedať hlbokým múdrym hlasom: „To je posvätné tajomstvo!“ Existujú otázky, ktoré sa nemožno pýtať, a odpovede, ktoré nemožno pripustiť, aby sme obránili lož. Preto sa neodpovedateľnosť začne spájať s posvätnosťou. A cenou za ochranu pred kritikou je vzdanie sa skutočnej zvedavosti, ktorá si naozaj želá nájsť odpovede. Budete uctievať svoju vlastnú nevedomosť ohľadom dočasne nezodpovedaných otázok vašej generácie – pravdepodobne zahŕňajúc aj tie, ktoré už sú zodpovedané.

Viera – v raných dňoch náboženstva, keď boli ľudia naivnejší, keď ešte aj inteligentní ľudia naozaj tomu všetkému verili, náboženstvá zakladali svoju povesť na svedectve o zázrakoch v ich písmach. A kresťanskí archeológovia sa vybrali pátrať, naozaj očakávajúc, že nájdu trosky Noemovej archy. Ale keď neprichádzali žiadne takéto indície, *potom* náboženstvo urobilo to, čo William Bartley nazval *únik do záväzku*: „Verím, pretože verím!“ Tak sa *viera bez dobrých indícií* začala spájať so zážitkom posvätného. A cenou za ochranu pred kritikou je, že obetujete svoju schopnosť jasne rozmýšľať o tom, čo je posvätné, a pokračovať vo svojom chápaní posvätného, a vzdávať sa chýb.

→ http://lesswrong.com/lw/oz/scarcity/

201 Adam Frank, *The Constant Fire: Beyond the Science vs. Religion Debate* (University of California Press, 2009).

Zážitkovosť – ak ste si predtým mysleli, že dúha je posvätná zmluva Boha s ľudstvom, a potom ste si začali uvedomovať, že Boh neexistuje, môžete urobiť *únik do číreho zážitku* – chváliť sami seba skrátka za to, že *cítite* také úžasné pocity, keď myslíte na Boha, bez ohľadu na to, či Boh naozaj *existuje* alebo nie. A cena za ochranu pred kritikou je solipsizmus: váš zážitok je zbavený *referentov*. Aký hrozne prázdny pocit by to bol sledovať, ako sa raketoplán dvíha na pilieri ohňa a hovoriť si: „Ale v skutočnosti nezáleží na tom, či ten raketoplán naozaj existuje alebo nie, hlavne že mám tento pocit."

Oddelenie – ak ríša posvätného nepodlieha bežným pravidlám indície, ani sa nedáskúmať bežnými prostriedkami, potom musí byť iného druhu než svet obyčajnej hmoty: a tak máme menší sklon považovať raketoplán za kandidáta na posvätnosť, pretože je to dielo iba *ľudských* rúk. Keats prestal obdivovať dúhu a degradoval ju do „nudného katalógu všedných vecí" za ten zločin, že bolo známe jej tkanivo a textúra. A cena za ochranu pred všednou kritikou je, že stratíte posvätnosť všetkých iba skutočných vecí.

Súkromie – o tom som už hovoril.

Takéto skreslenia sú dôvodom, prečo je lepšie *nepokúšať* sa zachraňovať náboženstvo. Nie, ani len ako formu „spirituality". Odoberte inštitúcie a faktické chyby, odčítajte kostoly a písma, a zostane vám... celý tento nezmysel o tajomnosti, viere, solipsistickom zážitku, súkromnej samote a nespojitosti.

Pôvodná lož je iba začiatkom tohto problému. Potom máte všetky tie zlozvyky myslenia, ktoré sa vyvinuli, aby ju bránili. Náboženstvo je otrávená čaša, z ktorej by bolo lepšie nevypiť ani glg. Duchovno je tá istá čaša, z ktorej bola pôvodná tabletka jedu vybraná a zostala iba tá rozpustená časť – o čosi menej priamo smrtiace, ale aj tak nie dobré.

Keď bola nejaká lož bránená celé veky, skutočný dôvod zdenených zvykov sa stratil v hmle, vrstva za vrstvou nedokumentovaných chorôb; potom si myslím, že múdri ľudia začnú znovu od začiatku, namiesto snahy selektívne odhodiť pôvodnú lož, ale uchovať si zvyky myslenia, ktoré ju chránili. *Skrátka si priznajte, že ste sa mýlili*, vzdajte sa svojej chyby *úplne*, prestaňte ju *hocijako* chrániť, prestaňte sa snažiť povedať, že ste mali čo len trochu pravdu, prestaňte sa snažiť zachovať si tvár, jednoducho povedzte „Joj!" a zahoďte to *celé* a začnite odznova.

Táto schopnosť – naozaj, *naozaj*, bez obrany priznať, že ste sa *celkom* mýlili – je to, prečo nábožensky zážitok nikdy nebude ako vedecký zážitok. Žiadne náboženstvo nedokáže prijať *túto* schopnosť bez toho, aby *celkom* nestratilo samo seba a nestalo sa jednoducho ľudstvom...

...pozerajúcim jednoducho nahor na vzdialené hviezdy. Uveriteľné bez námahy, bez ustavičného rozptyľujúceho úsilia odstrkovať svoje uvedomovanie si protiindície. Skutočne tam *vo svete*, zážitok zjednotený so svojím referentom, pevná časť toho odvíjajúceho sa príbehu. Poznateľné bez hrozby, ponúkajúce skutočnú potravu pre zvedavosť. Zdieľané spolu s mnohými ďalšími prizerajúcimi sa, bez potreby utekať do súkromia. Vytvorené z rovnakého tkaniva ako vy a ako všetky ostatné veci. To najsvätejšie a najkrajšie, posvätné obyčajno.

*→

212. Aby ste šírili vedu, držte ju v tajnosti

Občas rozmýšľam, či Pytagorovci nemali ten správny nápad.

Áno, písal som o tom, ako je „veda" nevyhnutne verejná. Písal som, že rozdiel medzi „vedou" a púhym rozumným poznaním je principiálna možnosť reprodukovať vedecké pokusy sám, bez nutnosti spoliehať sa na autoritu. Povedal som, že „vedu" by sme mali definovať ako verejne dostupné poznanie ľudstva. Dokonca som naznačil možnosť, že budúce generácie budú považovať všetky články neuverejnené v časopisoch s otvoreným prístupom za nie-vedu, čiže nemôže to byť súčasťou verejného poznania ľudstva, ak za prečítanie musia ľudia platiť.

Ale to je iba jedna predstava budúcnosti. V inej predstave sa vedomosti, ktoré dnes nazývame „veda" *odstránia* z verejnej oblasti – knihy a časopisy sa skryjú, budú ich chrániť mystické sekty guruov

→ http://lesswrong.com/lw/57/the_sacred_mundane/

oblečených v rúchach, vyžadujúce strašné iniciačné rituály pri vstupe – takže ich viac ľudí bude *naozaj* študovať.

Chcem tým povedať, že teraz ľudia *môžu* študovať vedu, ale *nerobia* to.

V sociálnej psychológii sa to volá „vzácnosť“. Čo sa zdá, že je v obmedzenom množstve, sa cení viac. A tento efekt je *mimoriadne* silný pri informáciách – máme omnoho väčší sklon pokúsiť sa získať informáciu, o ktorej veríme, že je tajná, a viac si ju vážiť, keď ju získame.

Pri vede sa mi zdá, že ľudia predpokladajú, že ak je informácia voľne dostupná, nemôže byť dôležitá. Takže ľudia namiesto toho vstupujú do siekt, ktoré majú dosť rozumu, aby svoje Veľké Pravdy držali v tajonosti. Tá Veľká Pravda môže byť v skutočnosti táranina, ale je omnoho uspokojujúcejšia než súvislá veda, pretože je *tajná*.

Veda je veľký Odcudzený List našej doby, ponechaný všetkým na očiach a ignorovaný.

Iste, vedecká otvorenosť pomáha vedeckej elite. Tí si už *prešli* cez iniciačné rituály. Ale pre zvyšok planéty je veda stonásobne účinnejšie utajená tým, že je voľne dostupná, než keby jej knihy boli strážené v trezoroch a museli by ste kráčať po žeravých uhlíkoch, aby ste získali prístup. (To je veru obávaná skúška, pretože veľké tajomstvá izolácie sú dostupné iba Zasväteným Fyzikom na Tretej Úrovni.)

Keby vedecké poznatky boli skryté v prastarých trezoroch (a nie iba skryté v nepohodlných časopisoch s plateným prístupom), ľudia by sa aspoň *snažili* dostať do tých trezorov. Zúfalo by sa *chceli* naučiť vedu. Najmä keby videli moc, ktorou vládnu Fyzici na Ôsmej Úrovni, a bolo by im povedané, že *nemajú právo poznať* vysvetlenie.

A keby ste sa pokúsili vytvoriť sektu okolo, povedzme, scientológie, spočiatku by ste dostali nejaký stupeň verejného záujmu. Ale ľudia by veľmi rýchlo začali klásť nepohodlné otázky ako: „Prečo verejne nepredvádzate schopnosti vašej Ôsmej Úrovne, tak ako Fyzici?“ a „Ako je možné, že žiaden z Majstrov Matematikov sa nechce pridať k vašej sekte?“ a „Prečo by so mal nasledovať vášho Zakladateľa, keď nie je na Ôsmej Úrovni nikde mimo svoje vlastnej sekty?“ a „Prečo by som mal študovať *vašu* sektu ako *prvú*, keď Zlovestní Zubári dokážu robiť veci, ktoré sú omnoho úžasnejšie?“

Keď sa na to pozriete z tejto perspektívy, únik matematiky z Pytagorejskej sekty začína vyzerať ako veľká strategická chyba ľudstva.

Viem, že sa teraz opýtate: „Ale veda *je* obklopená strašnými rituálmi zasvätenia! Navyše je *vo svojej podstate* náročná na učenie! Prečo sa *toto* nepočíta?“ Pretože verejnosť *si myslí*, že veda je voľne dostupná, preto. Ak máte *dovolené* sa ju učiť, nemôže byť dosť dôležitá na to, *aby* sa ju niekto učil.

Je to problém imidžu, ľudí riadiacich sa cudzími postojmi. *Hocikto* môže prísť do supermarketu a kúpiť si žiarovku, takže na ňu nikto nepozerá s úžasom a úctou. Jej fyzika údajne nie je tajomstvo (aj keď ju *vy* nepoznáte) a v novinách je vysvetlenie na jeden odstavec, ktoré znie hmlisto autoritatívne a presvedčivo – skrátka, nikto nepovažuje žiarovku za posvätné tajomstvo, preto ani vy.

Ešte aj tie najjednoduchšie maličkosti, celkom nereagujúce predmety ako krucifixy, sa môžu stať magickými, ak na ne každý *pozerá* akoby boli magické. Ale keďže máte teoreticky *dovolené* vedieť, prečo žiarovka funguje, aj bez vylezenia na horu a hľadania opusteného Kláštora Elektrikárov, nie je dôvod *naozaj* sa unúvať učiť.

Nuž, pretože veda v skutočnosti má rituály zasvätenia, aj spoločenské aj poznávacie, vedci nie sú so svojou vedou celkom nespokojní. Problém je, že v dnešnom svete sa veľmi málo ľudí vôbec unúva študovať vedu. Veda nemôže byť to pravé Tajné Poznanie, skrátka pretože má každý dovolené poznať ju – *hoci v skutočnosti ju nepoznajú*.

Keby sa vám Veľké Tajomstvo Prirodzeného Výberu, odovzdávané od Darwina Ktorý Nie Je Zabudnutý, sprístupnilo až po zaplatení 2000 dolárov a absolvovaní ceremónie zahŕňajúcej fakle a róby a masky a obetovanie býka; keby vám *potom* ukázali skameneliny, ukázali pod mikroskopom zrakový nerv vedúci cez sietnicu, a nakoniec povedali Pravdu, povedali by ste: „Toto je tá najúžasnejšia vec na svete!“ a *boli by ste spokojní*. A keby sa vám potom nejaká iná sekta snažila povedať, že to bol v skutočnosti bradatý muž na oblohe pred 6000 rokmi, smiali by ste sa o dušu.

A viete, možno by to naozaj mohlo byť takto *zábavnejšie*. Najmä keby zasvätenie vyžadovalo, aby ste sami dali dohromady nejaké indície – napríklad so spolužiakmi – skôr než poviete svojmu Senseiovi

Vedy, že ste pripravení pustúpiť na nasledujúcu úroveň. Nebolo by to *efektívne*, iste, ale bola by to *zábava*.

Keby ľudstvo nikdy nebolo urobilo tú chybu – keby nikdy nešlo náboženskou cestou, a nenaučilo sa báť všetkého, čo zaváňa náboženstvom – potom by možno obrad udeľovania Ph.D. zahŕňal litánie a spevy, pretože to je to, čo sa ľuďom páči. Prečo zo všetkého uberať zábavu?

Možno to robíme celé zle.

Ale nie, nenavrhujem tu *vážne*, aby sme sa pokúsili otočiť posledných päťsto rokov otvorenosti a klasifikovali všetku vedu ako tajnú. Prinajmenšom nie ihneď. Teraz je dôležitá efektívnosť, najmä vo veciach ako lekársky výskum. Iba vysvetľujem, prečo nechcem nikomu povedať Tajomstvo vzniku nevysloviteľného rozdielu medzi modrosťou a červenosťou z púhych atómov za menej ako 100 000 dolárov...

Ehm! Chcel som povedať, že vám hovorím o tejto vízii alternatívnej Zeme, aby ste dali vede rovnaké postavenie ako sektám. Aby ste nepodceňovali vedeckú pravdu, keď sa ju naučíte, *iba* preto, lebo nevyzerá chránená primerane jej cene. *Predstavte* si tie rúcha a masky. Predstavte si, ako sa zakrádate do trezorov a kradnete Newtonove Stratené Vedomosti. A nenechajte sa oklamať žiadnou organizáciou, ktorá *používa* rúcha a masky, pokiaľ vám neukážu aj svoje údaje.

Zdá sa, že ľudia majú v mysliach dieru pre Ezoterické Poznanie, Hlboké Tajomstvá, Skrytú Pravdu. A ja túto psychológiu ani nekritizujem! *Existujú* hlboké tajné ezoterické skryté pravdy, ako kvantová mechanika alebo bayesovská štruktúra. Len sme si privykli prezentovať túto Skrytú Pravdu veľmi *neuspokojivým* spôsobom, zabalenú do falošnej všednosti.

Lenže ak tie diery pre tajné poznanie nebudú zaplnené pravdivými názormi, budú zaplnené nepravdivými názormi. Nemáme sa učiť čo *iné ako* vedu – emocionálnu energiu musíme buď investovať do skutočnosti, alebo vyplytvať na úplný nezmysel, alebo zničiť. Ja si myslím, že je lepšie túto emocionálnu energiu investovať; netreba zábavu zbytočne odhadzovať.

Práve teraz máme to najhoršie z oboch svetov. Veda nie je *naozaj* voľne dostupná, pretože kurzy sú drahé a učebnice sú drahé. Ale verejnosť si *myslí*, že každý má právo to vedieť, preto to nemôže byť dôležité.

Ideálne by sme chceli, aby tieto veci fungovali presne naopak.

*→

213. Obrad zasvätenia

Fakle, ktoré ožarovali úzke schodisko, horeli intenzívne a nesprávnou farbou, plameň ako taviace sa zlato alebo roztrieštené slnká.

192... 193...

Brennanove sandále jemne ťukali o kamenné schody, postupne dopadajúc, ako keď veľmi pomaly padá domino.

227... 228...

Pol kruhu pred ním sa koniec temného plášťa sunul dole schodmi, postava v rúchu sa držala tesne mimo dosah.

239... 240...

Už to nebude dlho, predpovedal pre seba Brennan, a jeho odhad bol správny:

To číslo bolo šestnásťkrát šestnásť a stáli pred sklenou bránou.

Veľká zakrivená brána bola vypracovaná s rafinovanosťou, humorom a jemnou pozornosťou indexom lomu: zakrivovala svetlo, ohýbala ho, a všeobecne ho zneužívala, takže tu boli náznaky, čo je na druhej strane (silnejšie zdroje svetla, temné steny), ale žiaden spôsob ako *vidieť cez ňu* – pokiaľ ste samozrejme nemali kľúč: opačné dvere, hrubé oproti tenkému a tenké oproti hrubému, čím by sa obe navzájom zrušili.

→ http://lesswrong.com/lw/p0/to_spread_science_keep_it_secret/

Z postavy v rúchu vedľa Brennana sa vynorili dve ruky, pokryté odrážavou látkou, ktorá skrývala farbu pokožky. Prsty ako štíhle zrkadlá uchopili držadlá na zakrivenej bráne – držadlá, ktoré by Brennan nebol odhadol; pri všetkom tom skreslení sa tvary dali iba odhadovať, nie vidieť.

„Naozaj chceš vedieť?“ zašepkal sprievodca; šepot takmer rovnako hlasný ako bežný hlas, ale neprezrádzajúci najmenší náznak pohlavia.

Brennan zaváhal. Odpoveď na túto otázku znela podozrivo, priam mimoriadne samozrejme, ešte aj na rituál.

„Áno,“ povedal Brennan nakoniec.

Sprievodca naňho iba mlčky pozeral.

„Áno, chcem vedieť,“ povedal Brennan.

„Vedieť *čo*, presne?“ zašepkala postava.

Brennanova tvár sa zmraštila sústredením, pokúšal sa predstaviť si túto hru po jej koniec, a dúfal, že to už nepokazil; až sa nakoniec spoľahol na svoje prvé a posledné útočisko, ktorým bola pravda:

„To je jedno,“ povedal Brennan, „odpoveď je stále áno.“

Sklená brána sa v strede rozdelila a zasunula sa, iba s najtichším zakšrípaním, do okolitého kameňa.

Odhalená miestnosť bola lemovaná, od steny k stene, postavami v rúchach a kapucniach z látky pohlcujúcej svetlo. Samotné rovné steny neboli z čierneho kameňa, ale odrážali čierny štvorec temných rúch v mriežke do nekonečna na všetky strany; takže to vyzeralo, akoby ľudia z nejakého omnoho rozľahlejšieho mesta, alebo azda celé ľudstvo, sledovali v zhromaždení. Vo vzduchu bol náznak vlhkého tepla, dych zhromaždených: vôňa davu.

Brennanov sprievodca sa presunul do stredu štvorca, kde horeli štyri fakle toho neúprosného žltého plameňa. Brennan nasledoval, a keď zastal, uvedomil si s miernym šokom, že všetky tie skryté kapucne teraz pozerajú priamo naňho. Brennan predtým nikdy v živote nebol ohniskom takej absolútnej pozornosti; bolo to desivé, hoci nie celkom nepríjemné.

„Je tu,“ povedal sprievodca tým čudným hlasným šepotom.

Nekonečná mriežka postáv rúchach odpovedala jednohlasne: dokonale zmiešaní, presne zladení, takže sa žiaden jednotlivec nedal oddeliť od zvyšku, a povedali:

„*Kto chýba?*“

„Jakob Bernoulli,“ zaintonoval sprievodca a steny odpovedali:

„*Je mŕtvy, ale nie zabudnutý.*“

„Abraham de Moivre,“

„*Je mŕtvy, ale nie zabudnutý.*“

„Pierre-Simon Laplace,“

„*Je mŕtvy, ale nie zabudnutý.*“

„Edwin Thompson Jaynes,“

„*Je mŕtvy, ale nie zabudnutý.*“

„Zomreli,“ povedal sprievodca, „a stratili sme ich; ale stále máme jeden druhého a projekt pokračuje.“

V tichu sa sprievodca otočil k Brennanovi, a natiahol dlaň, na ktorej spočíval malý prsteň z takmer priesvitného materiálu.

Brennan vykročil, aby si vzal ten prsteň...

Ale ruka sa sa zavrela pevným stiskom.

„Ak tri štvrtiny ľudí v tejto miestnosti sú ženy,“ povedal sprievodca, „a tri štvrtiny žien a polovica mužov patria k Heréze Cnosti, a ja som Cnostný, aká je pravdepodobnosť, že som muž?“

„Dve jedenástiny,“ povedal Brennan sebavedome.

Chvíľu bolo absolútne ticho.

Potom šokované chichotanie.

Sprievodca opäť zašepkal, tentokrát naozaj ticho, takmer nečujne: „V skutočnosti je to jedna šestina.“

Brennanovi blčali líca tak, že si myslel, že sa mu celá tvár môže roztopiť. Cítil veľmi silný inštinkt vybehnúť z miestnosti, hore schodmi, ujsť do mesta, zmeniť si meno, začať svoj život odznova a tentokrát správne.

„Úprimná chyba je aspoň úprimná," povedal sprievodca, tentokrát hlasnejšie, „a úprimnosť poznáme podľa zriekania sa. Ak som Cnostný, aká je pravdepodobnosť, že som muž?"

„Jedna..." začal Brennan.

Potom sa zastavil. Opäť to hrozné ticho.

„Povedz už konečne ‚jedna šestina'," zašepkala postava, tentokrát dosť nahlas, aby to počuli aj steny; potom nasledoval ďalší smiech, nie vždy milý.

Brennan rýchlo dýchal a na čele mal pot. Ak sa v tomto pomýli, *naozaj* ujde z mesta. „Tri štvrtiny žien krát tri štvrtiny Cnostných, to je deväť šestnástin Cnostných žien v tejto miestnosti. Jedna štvrtina mužov krát jedna polovica Cnostných, to sú dve šestnástiny Cnostných mužov. Ak mám iba túto informáciu a fakt, že ste Cnostný, potom odhadujem šancu dve ku deviatim, čiže pravdepodobnosť dve jedenástiny, že ste muž. Hoci si v skutočnosti nemyslím, že daná informácia je správna. Po prvé, vyzerá to príliš upravene. Po druhé, v tejto miestnosti je nepárny počet ľudí."

Dlaň sa opäť vystrela a otvorila.

Brennan si vzal prsteň. Vo svetle fakieľ vyzeral takmer neviditeľný; nebol zo skla, ale z nejakého materiálu s indexom lomu podobným vzduchu. Prsteň bol teplý od ruky sprievodcu a zdal sa ako malá živá vec, keď sa ovinul okolo jeho prsta.

Úľava bola taká veľká, že takmer nepočul, ako skryté postavy tlieskajú.

Sprievodca naposledy zašepkal:

„Teraz si novicom Bayesovskej konšpirácie."

* →

→ http://lesswrong.com/lw/p1/initiation_ceremony/

R: Fyzikalizmus pre mierne pokročilých

214. Ruka verzus prsty

Späť k našej pôvodnej téme: Redukcionizmus a klam projekcie mysle. S prijatím redukcionizmu môžu byť emocionálne problémy, ak si myslíte, že veci musia byť nerozdeliteľné, aby mohli byť zábavné. Tento postoj nás však zaväzuje nikdy sa netešiť z ničoho zložitejšieho než kvark, takže ho radšej odmietam.

Aby sme si pripomenuli, redukcionistická téza je, že používame viacúrovňové modely kvôli výpočtom, ale fyzická skutočnosť má iba jednu úroveň.

Dnes by som rád položil nasledujúcu hádanku: Keď zdvihnete pohár vody, je to vaša *ruka*, čo ho zdvihlo?

Väčšina ľudí, samozrejme, súhlasí s naivnou obľúbenou odpoveďou: „Áno."

Nedávno však vedci urobili ohromujúci objav: Nie je to vaša *ruka*, čo drží pohár, v skutočnosti sú to vaše prsty a dlaň.

Áno, viem! Aj ja som bol šokovaný. Ale vyzerá to tak, že keď vedci odmerali sily, ktorými na pohár pôsobí každý z vašich prstov a vaša dlaň, zistili, že už nezvyšuje žiadna sila – takže sila, ktorou pôsobí vaša *ruka* musí byť nula.

Táto téma je tom, že ak dokážete *vidieť, ako* (nie iba *vedieť, že*) sa vyššia úroveň redukuje na nižšiu, nebudú vám pripadať ako oddelené veci na vašej mape; budete *vidieť*, aké hlúpe je myslieť si, že vaše prsty by mohli byť na jednom mieste a vaša ruka niekde inde; budete *vidieť*, aké pochabé je hádať sa o tom, či pohár dvíha vaša ruka alebo vaše prsty.

Kľúčové slovo je „vidieť", ako v konkrétnej predstave. Predstava vašej ruky spôsobuje, že si predstavujete prsty a dlaň; a naopak, predstava prstov a dlane spôsobuje, že v tomto myšlienkovom obraze identifikujete ruku. Takto sú vyššia úroveň *vašej mapy* a nižšia úroveň *vašej mapy* pevne spojené *vo vašej mysli*.

V skutočnosti sú samozrejme tieto úrovne spojené ešte pevnejšie – spojené tou najpevnejšiu možnou väzbou: fyzickou totožnosťou. Môžete to *uvidieť*: Môžete *uvidieť*, že povedanie (1) „ruka" alebo (2) „prsty a dlaň" neodkazuje na rôzne *veci*, ale na rôzne *uhly pohľadu*.

Predstavte si však, že by ste nemali vedomosť, aby ste tak pevne spojili úrovne svojej mapy. Napríklad by ste mohli mať „ručný skener", ktorý by ukazoval „ruku" ako bodku na mape (ako na starých radarových displejoch) a podobné skenery na prsty a dlane; potom by ste mohli vidieť zhluk bodiek okolo ruky, ale dokázali by ste si *predstaviť*, že sa bodka predstavujúca ruku pohne preč od ostatných. Takže aj keby fyzikálna skutočnosť ruky (čiže veci, ktorej táto bodka zodpovedá) bola totožná / striktne zložená z fyzikálnych skutočností prstov a dlane, nedokázali by ste vidieť tento fakt; aj keby vám to niekto povedal, alebo keby ste sa dovtípili zo súvislostí medzi bodkami, iba by ste tento fakt o redukcii *vedeli*, *nevideli* by ste ho. Stále by ste si dokázali *predstaviť* bodku pre ruku, ako sa hýbe nezávisle, hoci keby sa fyzické zloženie senzorov nemenilo, bolo by fyzikálne nemožné, aby sa to naozaj stalo.

Alebo, na ešte nižšej úrovni zviazania, by vám ľudia mohli skrátka povedať: „Tamto je ruka, a tamto sú prsty" - a v tom prípade by ste sotva vedeli viac než Stará Dobrá UI reprezentujúca celú situáciu pomocou sugestívne nazvaných symbolov LISP-u. Nebolo by nič *očividne* protikladné na tvrdení:

|— Vnútri(Izba, Ruka)

|— ~Vnútri(Izba, Prsty)

pretože by ste nemali *vedomosť*

|— Vnútri(x, Ruka) —> Vnútri(x, Prsty)

Nič z tohto nehovorí, že ruka v skutočnosti môže odpojiť svoju existenciu od vašich prstov a plaziť sa ako duch krížom bez miestnosť; akurát to hovorí, že Stará Dobrá UI s výrokovou reprezentáciou o tom nemusí *vedieť* nič viac. Mapa nie je územie.

Konkrétne by ste nemali odvodzovať priveľa záverov z toho, že v mysli nejakého konkrétneho mysliteľa vyzerá *pojmovo možné* oddeliť ruku od prvkov, z ktorých sa skladá, prstov a dlane. Pojmová možnosť nie je to isté ako logická možnosť alebo fyzikálna možnosť.

Pre vás je *pojmovo možné*, že 235757 je prvočíslo, pretože neviete nič viac. Ale nie je *logicky* možné, aby 235757 bolo prvočíslo; keby ste boli logicky vševediaci, 235757 by bolo samozrejme zložené číslo (a vedeli by ste jeho činitele). Preto máme pojem možných nemožných svetov, aby sme vedeli prideľovať pravdepodobnostné distribúcie výrokom, ktoré môžu ale nemusia byť *v skutočnosti* logicky nemožné.

A dokážete si predstaviť filozofov, ktorí kritizujú „eliminatívnych prstovcov“, ktorí odporujú priamym faktom zo skúsenosti – dokážeme predsa *cítiť*, že naša ruka drží pohár – tvrdením, že „ruky“ *v skutočnosti neexistujú*, lenže v tom prípade by pohár samozrejme spadol. A filozofov, ktorí navrhujú „zákony prepojenia prstov“ ako vysvetlenie ako konkrétna zostava prstov vyvoláva existenciu ruky – samozrejme s poznámkou, že hoci náš svet obsahuje tieto konkrétne zákony prepojenia prstov, tieto zákony mohli byť v princípe aj iné, takže to v nijakom zmysle nie sú *nevyhnutné fakty*, atď.

Všetko z tohto sú prípady klamu projekcie mysle, a to, čo nazývam „naivný filozofický realizmus“ - pomýlenie si filozofickej intuície s priamou, pravdivou informáciou o skutočnosti. Vaša neschopnosť predstaviť si niečo je iba výpočtový fakt o tom, čo si váš mozog vie a čo nevie predstaviť. Iný mozog by mohol fungovať odlišne.

*→

215. Zlostné atómy

Základná fyzika – kvarky a také veci – je veľmi ďaleko od úrovní, ktoré dokážeme *vidieť*, ako sú ruky a prsty. V najlepšom prípade môžete vedieť, ako replikovať pokusy, ktoré ukazujú, že vaša ruka (tak ako všetko ostatné) sa skladá z kvarkov, a môžete vedieť, ako odvodiť pár rovníc pre veci ako sú atómy a elektrónové oblaky a molekuly.

V najhoršom prípade je existencia kvarkov vo vašej ruke iba niečím, čo vám povedali. V tom prípade je pochybné, v akom vlastne zmysle možno povedať, že to všetko „viete“, dokonca aj keby ste zopakovali to isté slovo „kvark“, ktoré by fyzik použil na sprostredkovanie vedomosti inému fyzikovi.

V každom prípade nemôžete naozaj *vidieť* totožnosť medzi úrovňami – nikto nemá dosť veľký mozog na to, aby si *predstavil* avogadrove množstvá kvarkov a rozoznal v nich vzor ruky.

Ale aspoň rozumieme, čo ruky *robia*. Ruky tlačia veci, vyvíjajú na ne silu. Keď hovoríme o atómoch, predstavujeme si malé biliardové gule, ako do seba narážajú. Preto vyzerá zrejmé, že aj „atómy“ môžu tlačiť na veci tým, že do nich narážajú.

Táto predstava atómov však nie je celkom správna. Ale pokiaľ siaha *ľudská predstavivosť*, je pomerne ľahké predstaviť si, ako sa vaša ruka skladá z malej galaxie víriacich biliardových gúľ, ktoré tlačia na veci, keď sa ich vaše „prsty“ dotknú. Démokritos si to predstavoval pred 2400 rokmi, a bolo obdobie, približne 1803-1922, keď si Veda myslela, že mal pravdu.

Ale čo povedzme taký hnev?

Ako by sa malé biliardové gule mohli hnevať? Malé zamračené tváre na biliardových guliach?

Predstavte si sami seba na mieste napríklad lovca-zberača – niekoho, kto možno ani nemá pojem pre písanie, tobôž pojem pre používanie obyčajnej hmoty na vykonávanie výpočtov – niekoho, kto ani netuší, že existuje niečo také ako neuróny. Potom si môžete predstaviť tú *funkčnú* priepasť, ktorú vaši predkovia mohli vnímať medzi biliardovými guľami a „Grrr! Aaarg!“

Zabudnite na chvíľu na subjektívne skúsenosti a vezmite si tú šíru priepasť *v správaní* medzi hnevom a biliardovými loptami. Rozdiel medzi tým, čo *robia* malé biliardové gule, a čo hnev *robí* s ľuďmi. Hnev môže spôsobiť, že ľudia zdvihnú päsť a niekoho udrú – alebo povedia posmešné veci za jeho chrbtom – alebo mu v noci strčia do stanu škorpióna. Biliardové gule iba posúvajú veci.

→ http://lesswrong.com/lw/p2/hand_vs_fingers/

Skúste sa sami vžiť do role lovca-zberača, ktorý nikdy nemal „Aha!“ spracovania informácií. Skúste sa vyhnúť skresleniu spätného pohľadu ohľadom vecí ako sú neuróny a počítače. Iba potom dokážete uvidieť tú neprekročiteľnú priepasť vo vysvetľovaní:

Ako môžete vysvetliť zlostné správanie pomocou biliardových gúľ?

Nuž, *samozrejmý* materialistický odhad je, že tieto malé biliardové gule zatlačia na vaše rameno a spôsobia, že niekoho udriete, alebo zatlačia na váš jazyk tak, že z neho vyjdú urážky.

Ale ako tieto malé biliardové gule vedia, ako to majú urobiť – alebo ako dlhodobo navigovať váš jazyk a prsty – ak sami nie sú zlostné?

A okrem toho, ak vás nezlákal – ach! - scientizmus, môžete vidieť z pohľadu prvej osoby, že toto vysvetlenie je jasne zlé. Atómy môžu posunúť vaše rameno, ale nemôžu spôsobiť, že budete niečo *chcieť*.

Niekto by mohol podotknúť, že pitie vína vás môže urobiť zlostným. Ale kto tvrdí, že víno je urobené iba z malých biliardových gulí? Možno víno skrátka obsahuje potenciu hnevu.

Je zrejmé, že redukcionizmus je pomýlený postoj.

(Nováčik zablúdi a povie: „Moje umenie ma sklamalo“; majster zablúdi a povie: „Sklamal som svoje umenie.“)

Čo treba na prekročenie tejto priepasti? Nie je to iba myšlienka „neurónov“, ktoré „spracovávajú informácie“ - ak poviete iba toto a nič iné, iba tým vložíte magické nevysvetlené pravidlo na prechádzanie medzi úrovňami, kde prejdete z biliardu na myšlienky.

Ale programátor Umelej Inteligencie, ktorý vie, ako vytvoriť šachový program z obyčajnej hmoty, urobil *skutočný* krok k prekročeniu tejto priepasti. Ak rozumiete pojmom ako konsekvencializmus, spätné reťazenie, funkcie úžitku, a vyhľadávacie stromy, môžete vytvoriť čisto kauzálne / mechanické systémy, ktoré počítajú plány.

Ten trik funguje asi takto: Pre každý možný ťah v šachu vypočítajte ťahy, ktoré by mohol urobiť váš súper, potom vaše reakcie na tieto ťahy, a tak ďalej; vyhodnoťte čo najvzdialenejšiu viditeľnú pozíciu pomocou nejakého lokálneho algoritmu (môžete napríklad jednoducho spočítať figúrky); potom sa vráťte naspäť pomocou minimaxu a nájdite tak najlepší pohyb na aktuálnej šachovnici; potom urobte tento pohyb.

Všeobecnejšie: Ak máte v mysli reťaze kauzality, ktoré majú nejaké mapovanie – zrkadlia, odrážajú – čo sa deje v prostredí, potom môžete zbehnúť funkciu úžitku cez výsledné produkty predstavivosti, a nájsť akciu, ktorá dosiahne niečo, čo funkcia úžitku hodnotí vysoko, a vrátiť túto akciu. Nie je nevyhnutné, aby reťaze kauzality vo vašej mysli boli podobné prostrediu, aby sa skladali z biliardových gúľ, ktoré majú malé aury intencionality. Tranzistory počítača Deep Blue nemusia mať na sebe vyrezané malé šachové figúrky, aby fungovali. Pozrite si aj kapitolu Jednoduchá pravda.

Toto všetko je stále ohromne prehnane zjednodušené, ale malo by to prinajmenšom zmenšiť domnelú šírku tej priepasti. Ak dokážete pochopiť toto všetko, dokážete vidieť, ako plánovač postavený z obyčajnej hmoty môže byť ovplyvnený alkoholom, aby vyprodukoval viac zlostného správania. Biliardové gule v alkohole tlačia na biliardové gule, ktoré tvoria funkciu úžitku.

Ale aj keby ste vedeli ako písať malé UI, nedokážete si *predstaviť* prekročenie úrovní medzi tranzistormi a šachom. Je tam priveľa tranzistorov a priveľa ťahov, ktoré treba skontrolovať.

Podobne, aj keby ste vedeli všetky fakty o neurológii, nedokázali by ste si *predstaviť* prekročenie úrovní medzi neurónmi a hnevom – tobôž prekročenie úrovní medzi atómami a hnevom. Nie tak, ako si viete predstaviť ruku skladajúcu sa z prstov a dlane.

A čo ak vám kognitívny vedec len priamo povie: „Hnev sú hormóny“? Aj keď zopakujete tieto slová, neznamená to, že ste prekročili tú priepasť. Môžete veriť, že tomu veríte, ale to nie je to isté ako porozumenie, čo majú malé biliardové gule spoločné s chuťou niekoho udrieť.

Takže prídete s interpretáciami ako: „Hnev sú *iba* hormóny, je spôsobený malými molekulami, takže nemôže byť v žiadnom morálnom zmysle oprávnený – *preto* by ste sa mali naučiť ovládať svoj hnev.“

Alebo: „V skutočnosti neexistuje nič také ako hnev – je to ilúzia, citát bez referenta, ako fatamorgána vody na púšti, alebo hľadanie a nenájdenie draka v garáži.“

Toto sú obe tvrdé pilulky na prehltnutie (nie že by ste ich *mali* prehltnúť) a tak je omnoho ľahšie ich vyznávať, než v ne veriť.

Myslím si, že toto je to, čo si neredukcionisti / nematerialisti myslia, že kritizujú, keď kritizujú reduktívny materializmus.

Ale materializmus nie je taký ľahký. Nie je to také lacné ako povedať: „Hnev sa skladá z atómov – tak, hotovo.“ To by nevysvetlilo, ako sa dostať od biliardových gúľ k úderom. Potrebujete konkrétne vhľady do výpočtov, konsekvencializmu, a vyhľadávacích stromov, než začnete prekračovať vysvetľovaciu priepasť.

Toto všetko bol *na moderné pomery* pomerne ľahký príklad, pretože som sa sa obmedzil na hovorenie o zlostnom *správaní*. Hovoriť o výstupoch nevyžaduje ocenenie ako sa algoritmus cíti zvnútra (prekročenie priepasti medzi prvou a treťou osobou) alebo rozpustenie nesprávnej otázky (rozmotanie miest, kde vnútro vašej vlastnej mysle skresľuje skutočnosť).

Prejsť od hmotných látok, ktoré sa ohýbajú a lámu, horia a padajú, posúvajú a strkajú, k zlostnému *správaniu*, je iba cvičný problém podľa štandardov modernej filozofie. Ale je to *dôležitý* cvičný problém. Dokážeme ho naplno oceniť iba ak si uvedomíme, aké *ťažké* by bolo vyriešiť ho predtým než bolo objavené písmo. Raz tam bola vysvetľovacia priepasť – možno to tak nevyzerá v spätnom pohľade, keď už cez ňu stavali mosty celé generácie.

Vysvetľovacie priepasti môžu byť prekročené, ak prijmete pomoc od vedy a nebudete dôverovať pohľadu zvnútra svojej vlastnej mysle.

*→

216. Teplo verzus pohyb

Po predchádzajúcej kapitole mi napadlo, že existuje omnoho jednoduchší príklad redukcionistického preskočenia priepasti domnelého rozdielu druhu: redukcia z tepla na pohyb.

Dnes sa ekvivalencia tepla a pohybu môže zdať príliš zrejmou zo spätného pohľadu – každý hovorí, že „pohyb je teplo“, preto to nemôže byť „divný“ názor.

Ale boli časy, keď kinetická teória tepla bola vysoko kontroverznou vedeckou hypotézou, v kontraste k predstave kalorického toku, ktorý prúdil z teplejších objektov na chladnejšie. Ešte predtým bola hlavná teória tepla „Flogiston!“

Predstavte si, že by ste *oddelene* študovali kinetickú teóriu a kalorickú teóriu. Viete niečo o kinetike: zrážky, elastické odrazy, zotrvačnosť, kinetická energia, váha, zotrvačnosť, dráha voľného pádu. Oddelene viete niečo o teple: teploty, tlak, spaľovanie, tok tepla, stroje, topenie, vyparovanie.

Nielenže je takýto stav vedomostí uveriteľný, je to stav vedomostí, ktoré mal napríklad Sadi Carnot, ktorý pracujúc výhradne v rámci kalorickej teórie tepla vyvinul princíp Carnotovho cyklu – tepelný stroj s maximálnou účinnosťou, z ktorého existencie vyplýva druhý termodynamický zákon. To bolo v roku 1824, keď kinetika bola vysoko rozvinutou vedou.

Predstavte si, že ako Carnot viete veľa o kinetike, a veľa o teple, ako *samostatné* veci. Teda samostatné časti *poznania*: váš mozog má oddelené triediace košíky na názory o kinetike a názory o teple. Ale zvnútra takýto stav vedomostí *vyzerá* ako žiť vo svete pohyblivých vecí a horúcich vecí, svete, kde pohyb a teplota sú nezávislé vlastnosti hmoty.

Teraz príde Fyzik z Budúcnosti a povie vám: „Kde je teplo, tam je pohyb, a naopak. Preto sa napríklad veci trením o seba zohrievajú.“

Sú (prinajmenšom) dve možné vysvetlenia, ktoré by ste mohli pripísať výroku: „Kde je teplo, tam je pohyb, a naopak.“

Po prvé, mohli by ste predpokladať, že teplo a pohyb existujú samostatne – že kalorická teória je správna – ale že medzi fyzikálnymi zákonmi nášho vesmíru je „premosťovací zákon“, ktorý hovorí, že kde sa objekty pohybujú rýchlo, začnú vznikať kalórie. A naopak, iný premosťovací zákon hovorí, že

→ http://lesswrong.com/lw/p3/angry_atoms/

kalórie môžu vyvíjať tlak na veci a rozhýbať ich, čo je dôvod, prečo teplejší plyn vyvíja viac tlaku na svoj obal (a tak môže parný stroj použiť paru na pohyb piestu).

Po druhé, mohli by ste predpokladať, že teplo a pohyb sú, v nejakom zatiaľ neznámom zmysle, *tá istá vec*.

„Nezmysel," povie Mysliteľ 1, „slová ‚teplo' a ‚pohyb' majú dva odlišné významy; preto máme dve odlišné slová. Vieme, ako určiť, kedy budeme pozorovanú udalosť nazývať ‚teplo' - teplo môže roztaviť veci, alebo spôsobiť, že vzbĺknu plameňom. Vieme ako určiť, kedy povieme, že nejaký predmet sa ‚rýchlo pohybuje' - mení polohu; a keď narazí, môže sa pokrčiť alebo rozbiť. Teplo sa týka zmeny hmoty; pohyb sa týka zmeny polohy a tvaru. Povedať, že tieto dve slová majú rovnaký význam je iba zmätenie seba samého."

„Nemožné," povie Mysliteľ 2. „Možno sú v našom svete teplo a pohyb spojené premosťovacími zákonmi, takže je fyzikálny zákon, že pohyb vytvára kalórie, a naopak. Ale viem si ľahko predstaviť svet, kde sa veci trením o seba *nezohrievajú*, a kde plyny *nevyvíjajú* väčší tlak pri vyšších teplotách. Keďže existujú možné svety, kde teplo a pohyb nie sú spojené, musia to byť rôzne vlastnosti – toto je pravda a priori."

Mysliteľ 1 si pletie citát a referent. 2 + 2 = 4, ale „2 + 2" ≠ „4". Reťazec „2 + 2" obsahuje päť znakov (vrátane medzier) a reťazec „4" obsahuje iba 1 znak. Ak napíšete tieto dva reťazce do interpretera Pythonu, vypíšu rovnaký výsledok, 4. Nemôžete teda uzavrieť z pohľadu na reťazce „2 + 2" a „4", že len pretože sú tieto reťazce rôzne, musia mať rôzny „význam" pre interpreter Pythonu.

O slovách „teplo" a „kinetická energia" sa dá povedať, že „odkazujú na" tú istú vec, ešte predtým než *vieme*, že teplo sa redukuje na pohyb, v tom zmysle, že ešte nevieme, čo je táto referencia, ale že tie referencie sú v skutočnosti to isté. Mohli by ste si predstaviť Idealizovaný Vševediaci Interpreter Vedy, ktorý by dal rovnaký výstup, keď by sme napísali „teplo" a „kinetická energia" do príkazového riadku.

Hovorím o Interpreteri Vedy, aby som zdôraznil, že na dereferencovanie ukazovateľa musíte vystúpiť mimo poznania. Výsledok dereferencovania je niečo tam vonku v skutočnosti, nie v niekoho hlave. Môžete teda *povedať* „skutočný referent" alebo „aktuálny referent", ale nemôžete tieto slová *vyhodnotiť* lokálne, vo vnútri svojej vlastnej hlavy. Nemôžete uvažovať pomocou skutočného referentu tepla – keby vaše myšlienky používali *skutočné teplo*, myšlienka „1 milión kelvinov" by vyparila váš mozog. Vytvorením názoru o svojom názore o teple však môžete hovoriť *o* svojom názore o teple, a povedať veci ako: „Je možné, že môj názor o teple až tak nepripomína *skutočné* teplo." Nemôžete toto porovnanie naozaj urobiť priamo vo svojom mozgu, ale môžete *o ňom* hovoriť.

Môžete teda povedať: „Moje názory o teple a pohybe nie sú rovnaké názory, ale je možné, že skutočné teplo a skutočný pohyb sú tá istá vec." Je to ako byť schopný uznať, že „zornička" a „večernica" by mohli byť tá istá planéta, hoci zároveň rozumiete, že toto nedokážete určiť iba skúmaním svojich názorov – musíte vytiahnuť dalekohľad.

Chyba Mysliteľa 2 je podobná. Fyzik mu povedal: „Kde je teplo, tam je pohyb" a M2 to nesprávne pochopil ako výrok o *fyzikálnom zákone*: Prítomnosť kalórií *spôsobuje* existenciu pohybu. Zatiaľ čo fyzik naozaj myslel niečo skôr podobné *usudzovaciemu pravidlu*: Kde ti povedia, že je „teplo", usudzuj na prítomnosť „pohybu".

Z tejto základnej projekcie viacúrovňového modelu na viacúrovňovú skutočnosť vyplýva iná, ďalšia chyba: zjednotenie pojmovej možnosti s logickou možnosťou. Sadi Carnot si vie *predstaviť*, že by existoval iný svet, kde teplo a pohyb nesúvisia. Richard Feynman, ozbrojený konkrétnymi vedomosťami, ako odvodzovať rovnice o teple z rovníc o pohybe, si túto myšlienku nielenže nevie predstaviť, ale je natoľko nekonzistentná, že by mu z toho vybuchla hlava.

V rámci férovosti k filozofom musím poznamenať, že existujú filozofi, ktorí toto povedali. Napríklad Hilary Putnamová napísala o pokuse „dvojčaťa Zeme":[202]

202 Hilary Putnam, „The Meaning of Meaning," in *The Twin Earth Chronicles*, ed. Andrew Pessin and Sanford Goldberg (M. E. Sharpe, Inc., 1996), 3–52.

Keď raz objavíme, že voda (v skutočnom svete) je H2O, *nič sa nepočíta za skutočný svet, kde voda nie je H2O*. Konkrétne, ak za „logicky možný“ výrok označujeme taký, ktorý platí v nejakom „logicky možnom svete“, *nie je logicky možné, že by voda nebola H2O.*

Na druhej strane, vieme si dokonale predstaviť skúsenosti, ktoré by nás presvedčili (a preto by bolo rozumné veriť tomu), že voda *nie je* H2O. V tom zmysle je predstaviteľné, že voda nie je H2O. Je to predstaviteľné, ale nie je to logicky možné! Predstaviteľnosť nie je dôkazom logickej možnosti.

Zdá sa mi, že slovo „voda“ sa v týchto dvoch odsekoch používa v dvoch rôznych významoch – v jednom slovo „voda“ *odkazuje* na to, čo zadáme do Interpretera Vedy, a v jednom slovo „voda“ *odkazuje* na to, čo dostaneme z Interpretera Vedy, keď doňho zadáme „voda“. V prvom odseku Hilary asi hovorí, že keď urobíme nejaké pokusy a zistíme, že voda je H2O, voda sa automaticky predefinuje tak, že *znamená* H2O. Ale mohli by ste koherentne zastávať rôzne postoje ohľadom toho, či teraz slovo „voda“ *znamená* „H2O“ alebo „niečo, čo je *naozaj* vo fľaši vedľa mňa“, dokiaľ svoj pojem používate konzistentne.

Myslím si, že aj toto už bolo povedané. Každopádne...

Je celkom možné, že na svete existuje iba *jedna* vec, ale že nadobúda dostatočne odlišné podoby, a že vy ste dostatočne nevedomý ohľadom redukcie, takže vám to pripadá, že žijete vo svete obsahujúcom dve celkom rôzne veci. Vedomosti týkajúce sa týchto dvoch rôznych javov sa môžu učiť v rôznych triedach, a študovať v rôznych akademických oblastiach, ktoré sa nachádzajú v dvoch rôznych budovách vašej univerzity.

Musíte sa vrátiť ďaleko dozadu, do historicky realistického stavu mysle, aby ste si pamätali, aké *rôzne* sa teplo a pohyb kedysi zdali. Hoci, podľa toho, koľko veľa viete dnes, možno by to nebolo až také ťažké, keby ste dokázali pozrieť za tlak konvencie (že „teplo je pohyb“ je obyčajný názor a „teplo nie je pohyb“ je čudný názor). Myslím tým, predstavte si, že by zajtra fyzici vystúpili a povedali: „Naša popularizácia vedy vždy obsahovala jednu lož. V skutočnosti teplo nemá nič spoločné s pohybom.“ Vedeli by ste *dokázať*, že sa mýlia?

Povedať: „Možno sú teplo a pohyb tá istá vec!“ je ľahké. Ťažká časť je vysvetliť, *ako*. Vyžaduje to veľa podrobných vedomostí, aby ste sa dostali do bodu, kde už si nedokážete *predstaviť* svet, v ktorom by tieto dva javy išli každý svojou cestou. Redukcia nie je lacná, a preto si za ňu kúpite tak veľa.

Alebo by ste mohli povedať: „Redukcionizmus je ľahký, ale redukcia je ťažká.“ Ale myslím si, že trochu pomáha byť redukcionistom, keď príde čas hľadať redukciu.

* →

—

217. Prevratný objav mozgu! Skladá sa z neurónov!

Po úžasnom prevratnom objave medzinárodný tím vedcov na čele s laureátom Nobelovej ceny Santiagom Ramónom y Cajal oznámil, že mozog sa skladá z *absurdne* zložitej siete malých buniek, ktoré sú navzájom spojené mikroskopickými vláknami a vetvami.

Medzinárodný tím – ktorého súčasťou je aj známy technik Antonie van Leeuwenhoek, a možno aj Imhotep, vymenovaný za egyptského boha medicíny – vydal toto prehlásenie:

„Súčasný objav je vrcholom rokov výskumu naznačujúceho, že tá komplikovaná mäkká vec v našej lebke je zložitejšia než vyzerá. Vďaka Cajalovmu použitiu novej farbiacej techniky, ktorú vynašiel Camillo Golgi, sme sa dozvedeli, že táto štruktúra nie je spojitou sieťou ako sú cievy v tele, ale v skutočnosti sa skladá z mnohých maličkých buniek, alebo ‚neurónov‘, spojených navzájom ešte menšími vláknami.“

„Ďalšie rozsiahle indície, počínajúc gréckym lekárskym výskumníkom Alkmaiónom a pokračujúc výskumom rečových porúch Paula Brocu naznačujú, že mozog je sídlom rozumu.“

→ http://lesswrong.com/lw/p4/heat_vs_motion/

„Nemesius, biskup z Emesie, v minulosti tvrdil, že mozgové tkanivo je príliš pozemské, takže nemôže fungovať ako prostredník medzi telom a dušou, preto sa myšlienkové schopnosti nachádzajú v komorách mozgu. Keby však toto bolo správne, neexistoval by dôvod, prečo by tento orgán mal mať takú nesmierne zložitú vnútornú štruktúru.“

„Charles Babbage nezávisle naznačil, že mnohé malé mechanické prístroje by sa dali spojiť do *analytického stroja* schopného vykonávania aktivít ako je aritmetika, o ktorých sa všeobecne verí, že vyžadujú myslenie. Dielo Luigiho Galvaniho a Hermanna von Helmholtz naznačuje, že aktivity neurónov sú v podstate elektrochemické, a nie mechanické tlaky, ako sa verilo predtým. Napriek tomu si myslíme, že analógia s Babbageovým *analytickým strojom* naznačuje, že by mimoriadne komplikovaná sieť neurónov mohla podobne vykazovať myšlienkové vlastnosti.“

„Našli sme mimoriadne zložitý hmotný systém umiestnený tam, kde by mala byť myseľ. Dôsledky sú šokujúce a musíme im čeliť priamo. Veríme, že súčasný výskum ponúka silné experimentálne indície, že Benedictus Spinoza mal pravdu, a René Descartes sa mýlil: Myseľ a telo sú jednej podstaty.“

„V kombinácii s dielom Charlesa Darwina ukazujúcim, ako takýto zložitý orgán mohol v princípe vzniknúť v dôsledku procesov, ktoré samotné neboli inteligentné, väčšina vedeckých indícií teraz zrejme naznačuje, že inteligencia nie je ontologicky fundamentálna, ale postupne vznikala v čase. Toto je silný argument proti teóriám, ktorá pripisujú myšlienkovým entitám ontologicky fundamentálne alebo kauzálne prvotné postavenie, vrátane všetkých doteraz vynájdených náboženstiev.“

„Zostáva ešte mnoho práce pri objavovaní konkrétnych totožností medzi elektrochemickými reakciami medzi neurónmi, a myšlienkami. Napriek tomu veríme, že náš objav dáva prísľub, hoci ešte nie realizáciu, plného vedeckého vysvetlenia myslenia. Dnes môžeme tento problém vyhlásiť, ak nie za vyriešený, prinajmenšom za riešiteľný.“

Ľutujeme, že Cajal a väčšina ďalších výskumníkov v tomto projekte už nemohli poskytnúť komentár.

*→

218. Kedy sa antropomorfizmus stal hlúpym

Ukazuje sa, že väčšina vecí vo vesmíre nemá myseľ.

Toto tvrdenie by medzi mnohými dávnejšími kultúrami vyvolalo nedôverčivosť. Zvyčajný pojem je „animizmus“. Mysleli si, že stromy, kamene, rieky a kopce všetky mali duchov, pretože, hej, prečo nie?

Myslím tým, ak tie hrudy mäsa zvané „ľudia“ obsahujú myšlienky, prečo by nemohli aj hrudy dreva nazývané „stromy“?

Moje svaly sa hýbu podľa mojej vôle, a v rieke tečie voda. Kto môže povedať, že rieka nemá vôľu, ktorou hýbe vodu? Rieka pretiekla svoje brehy a zaplavila zhromaždisko mojej tlupy – prečo by som si nemal myslieť, že tá rieka sa hnevá, keď pohla svojou časťou, aby nám ublížila? To isté by som si myslel, keby niekoho päsť narazila do môjho nosa.

Neexistuje žiaden zrejmý dôvod – žiaden dôvod zrejmý *lovcovi-zberačovi* – prečo by to nemohlo byť tak. Vyzerá to ako *hlúpa* chyba iba ak si mýlite zvláštnosť s hlúposťou. Prirodzene, že nám viera, že rieky majú oživujúcich duchov, pripadá „zvláštna“, pretože to nie je viera našej tlupy. Ale nie je nič očividne hlúpe na myšlienke, že veľké hrudy pohybujúcej sa vody majú svojich duchov, tak ako naše vlastné hrudy pohybujúceho sa mäsa.

Keby tá myšlienka bola *samozrejme* hlúpa, nikto by jej nebol veril. Rovnako ako väčšinu času nikto neveril tej samozrejme hlúpej myšlienke, že Zem sa hýbe, keď vyzerá nehybne.

Je to také samozrejmé, že stromy nemôžu myslieť? Nezabúdajme na to, že stromy *sú v skutočnosti* našimi vzdialenými príbuznými. Vráťte sa dosť ďaleko dozadu a máte spoločného predka s papradím. Ak môžu hrudy mäsa myslieť, prečo nie aj hrudy dreva?

→ http://lesswrong.com/lw/p5/brain_breakthrough_its_made_of_neurons/

Aby bolo *samozrejmé*, že drevo nemyslí, musíte patriť do kultúry s mikroskopmi. Nie *hocijakými* mikroskopmi, ale naozaj *dobrými* mikroskopmi.

Aristoteles si myslel, že mozog je orgán na chladenie krvi. (Je dobré, že čo veríme o našich mozgoch, má veľmi malý účinok na ich skutočné fungovanie.)

Egypťania odhadzovali mozog počas procesu mumifikácie.

Alkmaión z Krotónu, Pytagorejec z 5. storočia pred naším letopočtom, určil mozog ako sídlo inteligencie, pretože vysledoval zrakový nerv z oka do mozgu. Aj tak, pri množstve indície, ktoré mal, to bol iba odhad.

Keby ústredná rola mozgu prestala byť odhadom? Neviem dosť z histórie na to, aby som odpovedal na túto otázku, a je pravdepodobné, že tam žiadna ostrá deliaca čiara nebola. Azda by sme ju mohli určiť v bode, kde niekto vysledoval anatómiu nervov a objavil, že prerušenie nervového spojenia *s mozgom* zablokuje pohyb a vnímanie?

Ešte aj tak máme iba tajomného ducha, ktorý sa hýbe cez nervy. Kto môže povedať, že drevo a voda, hoci im chýbajú tieto malé vlákna objavené v ľudskej anatómii, nemôžu prenášať toho istého tajomného ducha iným spôsobom?

Strávil som nejaký čas na internete v snahe nájsť tú presnú chvíľu, keď si niekto všimol tú vysoko spletitú vnútornú štruktúru neurónov v mozgu a povedal: „Hej, stavím sa, že celá táto obrovská motanica robí zložité spracovanie informácií!“ Nemal som veľa šťastia. (Nebol to Camillo Golgi – zamotanosť týchto obvodov bola známa pred Golgim.) Možno ani tu nikdy neexistoval žiaden dramatický okamih.

Ale tento objav tejto zamotanosti, a teória Charlesa Darwina o prirodzenom výbere, a predstava poznania ako výpočtu, je to, kde by som určil postupný začiatok úpadku antropomorfizmu do *samozrejmej* nesprávnosti.

Je to bod, v ktorom sa môžete pozrieť na strom a povedať: „Nevidím v biológii stromu nič, čo by robilo zložité spracovanie informácií. Ani to nevidím v jeho správaní, a ak je to skryté spôsobom, ktorý neovplyvňuje správanie stromu, ako by mohol vzniknúť selekčný tlak na takéto zložité spracovanie informácií?“

Je to bod, v ktorom sa môžete pozrieť na rieku a povedať: „Voda neobsahuje vzory replikujúce sa s dlhodobou dedičnosťou a podstatnou variáciou podliehajúcou iteratívnemu výberu, ako by teda rieka mohla mať nejaký vzor taký zložitý a funkčne optimalizovaný ako mozog?“

Je to bod, v ktorom sa môžete pozrieť na atóm a povedať: „Hnev možno vyzerá jednoducho, ale nie je jednoduchý, a nie je miesto, aby sa zmestil do niečoho takéto jednoduchého ako je atóm - pokiaľ teda vo vnútri kvarkov nie sú celé vesmíry podčastíc; a ešte aj tak, keďže sme nikdy nevideli žiaden príznak atómového hnevu, nemohol by mať žiaden účinok na javy na vyššej úrovni, ktoré poznáme.“

Je to bod, v ktorom sa môžete pozrieť na šteniatko a povedať: „Rodičia toto šteniatko mohli pritlačiť k zemi, keď urobilo niečo nesprávne, ale to neznamená, že toto šteniatko robí morálne uvažovanie. Naše súčasné teórie evolučnej psychológie tvrdia, že morálne uvažovanie vzniklo ako reakcia na zložitejšie spoločenské úlohy než je toto – naše morálne adaptácie vo svojej plno rozvinutej ľudskej forme sú výsledkom selekčných tlakov cez jazykové argumenty o kmeňovej politike.“

Je to bod, v ktorom sa môžete pozrieť na kameň a povedať: „Toto nemá ani jednoduchý vyhľadávací strom, aký je v šachovom programe – kde by to vzalo *úmysel* kotúľať sa dole kopcom, ako si kedysi myslel Aristoteles?“

Je napísané:

Čuang-c’ a Hui-Ši sa prechádzali pozdĺž priehrady vodopádu v Hao, keď Čuang-c’ povedal: „Pozri, ako tie rybky vyskakujú z vody od radosti! Toto sa rybám naozaj páči!“

Hui-Ši povedal: „Ty nie si ryba – ako môžeš vedieť, čo sa rybe páči?“

Čuang-c’ povedal: „Ty nie si ja, tak ako môžeš vedieť, že ja neviem, čo sa rybe páči?“

Teraz to už vieme.

*→

219. A priori

Tradičná Rozumnosť je formulovaná ako spoločenské pravidlá, ktorých porušenie sa interpretuje ako podvádzanie: ak porušíte nejaké pravidlo a nikto iný nerobí to isté, ste prvý porušovateľ – čiže zlý, zlý človek. Pre Bayesovcov je mozog strojom na presnosť: ak porušíte pravidlá rozumnosti, váš stroj nebude fungovať, a to platí bez ohľadu na to, či niekto iný porušuje pravidlá alebo nie.

Vezmime si problém Occamovej britvy, ako k nemu pristupujú tradiční filozofi. Ak sú dve hypotézy v rovnakom súlade s pozorovaniami, prečo veriť, že tá jednoduchšia je s väčšou pravdepodobnosťou správna? Mohli by ste argumentovať, že Occamova britva fungovala v minulosti, a preto je pravdepodobné, že bude fungovať v budúcnosti. Lenže toto samotné sa odvoláva na predpoveď podľa Occamovej britvy. „Occamova britva bude fungovať do 8. októbra 2007 a potom prestane fungovať“ je zložitejšie, ale je to v rovnakom súlade s pozorovanými indíciami.

Môžete argumentovať, že Occamova britva je rozumná distribúcia prvotných pravdepodobností. Ale čo je to „rozumná“ distribúcia? Prečo neoznačiť za „rozumnú“ veľmi komplikovanú prvotnú distribúciu, z ktorej by vyplývalo, že Occamova britva fungovala vo všetkých doteraz pozorovaných testoch, ale bude mať výnimky v budúcich?

Veru, zdá sa, že neexistuje spôsob, ako *obhájiť* Occamovu britvu, jedine *odvolaním* sa na Occamovu britvu, vďaka čomu tento *argument* asi *nepresvedčí* žiadneho *sudcu*, ktorý by Occamovu britvu už *neakceptoval*. (Čo je zvláštne na slovách napísaných kurzívou?)

Ak ste filozof, ktorého dennou prácou je písať články, kritizovať články druhých ľudí, a reagovať na cudziu kritiku vašich vlastných článkov, môžete sa pozrieť na Occamovu britvu a pokrčiť plecami. Tu je koniec obhajovania, argumentovania a presviedčania. Rozhodnete sa uzavrieť prímerie ohľadom písania článkov; ak vaši kolegovia filozofi nebudú žiadať zdôvodnenie vašich nevyargumentovateľných názorov, vy nebudete žiadať zdôvodnenie ich názorov. Ako symbol vašej dohody, vašu bielu vlajku, použijete frázu „a priori pravda“.

Avšak Bayesovcovi v tejto dobe kognitívnej vedy a evolučnej biológie a umelej inteligencie povedanie „a priori“ nevysvetlí, prečo mozgový stroj funguje. Ak má mozog nejakú úžasnú „továreň na apriórne pravdy“, ktorá *dokáže* vytvoriť presné názory, môžete sa čudovať, prečo smädní lovci a zberači nemohli použiť túto „továreň na apriórne pravdy“ na nájdenie pitnej vody. Môžete sa čudovať, prečo sa vôbec vyvinuli oči, ak existujú spôsoby vytvárania presných názorov aj bez pozerania sa na veci.

James R. Newman povedal: „Skutočnosť, že jedno jablko pridané k druhému jablku nemenne dáva dve jablká, pomáha pri vyučovaní aritmetiky, ale nemá nijaký vplyv na pravdivosť výroku, že 1 + 1 = 2.“ Internetová encyklopédia filozofie definuje „a priori“ tvrdenia ako tie, ktoré možno poznať nezávisle od skúsenosti. Wikipédie cituje Huma: Vzťahy medzi ideami sa „dajú objaviť pomocou púhej myšlienkovej operácie, nezávisle na tom, čo kde vo vesmíre existuje.“ Môžete vidieť, že 1 + 1 = 2 *samotným myslením*, bez pozerania na jablká.

Lenže v tejto dobe neurológie by si človek mal uvedomovať, že *myšlienky* existujú vo vesmíre; sú totožné s fungovaním mozgov. Hmotných mozgov, skutočných vo vesmíre, zložených z kvarkov v jednej jednotnej matematickej fyzike, ktorej zákony nerobia hranicu medzi vnútrom a vonkajškom vašej lebky.

Keď pomocou myslenia sčítate 1 + 1 a dostanete 2, samotné tieto myšlienky sú stelesnené v zábleskoch nervových vzorov. V princípe by sme mohli *pozorovať* rovnaké hmotné udalosti, ako sa dejú v mozgu niekoho iného. Vyžadovalo by si to nejaký pokrok vo výpočtovej neurobiológii a rozhraniach medzi mozgom a počítačom, ale v princípe by sa to dalo urobiť. Mohli by ste vidieť, ako stroj niekoho iného funguje hmotne, pomocou hmotných reťazí príčiny a následku, aby vypočítal „čistým myslením“, že 1 + 1 = 2. Ako sa pozorovanie tohto vzoru v mozgu *niekoho iného* líši, ako spôsob poznania, od pozorovania ako váš vlastný mozog robí to isté? Keď vám „čisté myslenie“ hovorí, že 1 + 1

→ http://lesswrong.com/lw/t5/when_anthropomorphism_became_stupid/

= 2 „nezávisle na nejakej skúsenosti či pozorovaní“, pozorujete tým vlastne svoj vlastný mozog ako indíciu.

Ak vám toto prípadá kontraintuitívne, skúste vidieť mysle/mozgy ako stroje – ako stroj, ktorý spojí nervový vzor pre 1 a nervový vzor pre 1 a dostane nervový vzor pre 2. Ak takýto stroj vôbec funguje, potom by mal *dávať rovnaký výstup*, keď pozoruje (pomocou očí a sietnice) podobný mozgový stroj vykonávajúci podobné spojenie a kopíruje do seba výsledný vzor. Inými slovami, pri každom druhu apriórneho poznania získaného „čistým myslením“ sa učíte presne to, čo by ste sa naučili, keby ste videli vonkajší mozgový stroj vykonávajúci tie isté čisté záblesky nervovej aktivácie. Tie stroje sú ekvivalentné, výstupy na spodnom riadku sú ekvivalentné, previazanosť názorov je rovnaká.

Nie je nič, čo viete „a priori“, čo by ste nemohli rovnako dobre vedieť pozorovaním chemických prenosov neurotransmiterov v nejakom vonkajšom mozgu. Čo si myslíte, že *ste*, drahý čitateľ?

To je dôvod, *prečo* viete predpovedať výsledok sčítania 1 jablka a 1 jablka tým, že si to najprv predstavíte v mysli, alebo naťukáte „3 × 4“ do kalkulačky, aby ste predpovedali výsledok predstavenia si 4 riadkov po 3 jablká v každom. Vy a jablko existujete v neohraničenom zjednotenom fyzikálnom procese, ktorého jedna časť môže odrážať druhú.

Sú tieto druhy nervových záblesk ov, ktoré filozofi označujú ako „apriórne pravdy“, *ľubovoľné*? Mnohé algoritmy UI fungujú lepšie s „regularizáciou“, ktorá skresľuje priestor riešení smerom k jednoduchším riešeniam. Ale samotné regularizované algoritmy sú zložitejšie; obsahujú riadok programu navyše (alebo 1000 riadkov navyše) v porovnaní s neregularizovanými algoritmami. Ľudský mozog je skreslený smerom k jednoduchosti, a preto myslíme efektívnejšie. Ak na tomto mieste stlačíte tlačidlo Ignoruj, zostal vám zložitý mozog, ktorý existuje bezdôvodne a funguje bezdôvodne. Neskúšajte mi teda nahovoriť, že „apriórne“ presvedčenia sú ľubovoľné, pretože určite neboli vytvorené hádzaním náhodných čísel. (Mimochodom, čo vlastne *znamená* prídavné meno „ľubovoľný“?)

Nemôžete sa vyhovoriť z vášho označovania výroku za „apriórny“ poukázaním na to, že aj *iní* filozofi majú problém zdôvodniť *svoje* výroky. Ak sa filozofovi niečo nepodarí vysvetliť, tento fakt nedokáže dodať elektrinu chladničke, ani fungovať ako kúzelná továreň na presné názory. Žiadne prímerie, žiadna biela vlajka, dokiaľ nepochopíte, prečo ten stroj funguje.

Ak zbavíte svoju myseľ *zdôvodňovania, argumentovania*, potom sa zdá zrejmé, prečo Occamova britva v praxi funguje: žijeme v jednoduchom svete, vo vesmíre s nízkou entropiou, v ktorom sa vyskytujú krátke vysvetlenia. „Ale,“ kričíte, „prečo je samotný vesmír pravidelný?“ Nuž to neviem, ale vidím to ako ďalšiu záhadu, ktorú bude treba vysvetliť. To nie je rovnaká otázka ako: „Ako budem argumentovať v prospech Occamovej britvy hypotetickému diskutérovi, ktorú ju ešte neakceptuje?“

Možnože nedokážete hypotetickému diskutérovi, ktorý neakceptuje Occamovu britvu, vyargumentovať vôbec *nič*; rovnako ako nedokážete nič vyargumentovať kameňu. Myseľ potrebuje mať určité množstvo dynamickej štruktúry, aby bola akceptovačom argumentov. Ak nejaká myseľ neimplementuje modus ponens, môže akceptovať „A“ aj „A -> B“ po celý deň a nikdy z toho neodvodiť „B“. Ako zdôvodníte modus ponens mysli, ktorá ho neakceptuje? Ako budete argumentovať kameňu, aby sa stal mysľou?

Mozgy sa vyvinuli z nemozgovej hmoty pomocou prirodzeného výberu; neboli vyargumentované do existencie hádaním sa s ideálnym študentom filozofie dokonalej prázdnoty. To neznamená, že naše úsudky sú nezmyselné. Mozgový stroj môže fungovať správne, produkovať správne názory, aj keď bol iba *postavený* – ľudskými rukami alebo kumulatívnymi stochastickými selekčnými tlakmi – namiesto vyargumentovania do existencie. Ale aby sa človek uspokojil s touto odpoveďou, musí vidieť rozumnosť v pojmoch strojov, nie argumentov.

*→

→ http://lesswrong.com/lw/k2/a_priori/

220. Reduktívna referencia

Redukcionistická téza (ako ju ja formulujem) je, že ľudské mysle z dôvodov efektivity používajú viacúrovňovú mapu, v ktorej *rozmýšľame* oddelene o veciach ako sú „atómy“ a „kvarky“, „ruky“ a „prsty“, alebo „teplo“ a „kinetická energia“. Samotná skutočnosť, na druhej strane, je jednoúrovňová v tom zmysle, že zrejme neobsahuje *samostatné, dodatočné, kauzálne pôsobiace* prvky na úrovni *nad* kvarkami.

Sadi Carnot formuloval druhý zákon termodynamiky (jeho predchodcu) pomocou kalorickej teórie tepla, v ktorej teplo bolo iba prúdenie z teplých vecí na studené veci, vytvárané ohňom, expandujúce plyny – účinky tepla sa študovali oddelene od vedy o kinetike, predtým než prišla redukcia. Ak sa pokúšate navrhnúť parný stroj, účinky všetkých tých maličkých vibrácií a zrážok, ktoré nazývame „teplo“, možno zhrnúť do omnoho jednoduchšieho popisu než je celá kvantová mechanika kvarkov. Ľudia počítajú efektívne, mysliac iba na významné účinky hodnôt súvisiacich s cieľom.

Ale samotná skutočnosť zrejme používa celú kvantovú mechaniku kvarkov. Raz som stretol chlapíka, ktorý veril, že keby ste použili všeobecnú relativitu na výpočet problému s nízkou rýchlosťou, ako delostrelecký granát, VR by vám dala *nesprávnu odpoveď* – nie iba pomalú odpoveď, ale *experimentálne nesprávnu odpoveď* – pretože pri nízkych rýchlostiach sa delostrelecké granáty riadia Newtonovskou mechanikou, nie VR. Práve takto fyzika *nefunguje*. Skutočnosť zrejme pokračuje v počítaní pomocou všeobecnej relativity aj tam, kde to robí rozdiel iba na štrnástom desatinnom mieste, čo by človek považoval za ohromné plytvanie výpočtovou silou. Fyzika to robí hrubou silou. Nikto *nikdy* neprichytil fyziku pri zjednodušovaní výpočtov – alebo ak áno, vládcovia Matrixu mu potom vymazali pamäť.

Naša mapa je teda veľmi nepodobná územiu; naše mapy sú viacúrovňové, územie je jednoúrovňové. Keďže táto reprezentácia je neuveriteľne nepodobná referentu, v akom zmysle môže byť názor ako „nosím ponožky“ považovaný za *pravdivý*, ak v samotnej skutočnosti existujú iba kvarky?

Ak ste zabudli, čo znamená slovo „pravdivý“, klasickú definíciu dal Alfred Tarski:

> Veta „sneh je biely“ je *pravdivá* vtedy a iba vtedy, keď sneh je biely.

V prípade, že ste zabudli, aký je rozdiel medzi vetami „Myslím si ‚sneh je biely‘“ a „‚Sneh je biely‘ je pravda“, pozrite tu. Pravdu nemožno vytvátiť *iba* tým, že sa pozriete dovnútra svojej hlavy – ak chcete napríklad vedieť, či „zornička = večernica“, potrebujete ďalekohľad; nestačí iba pozrieť na samotné názory.

Toto je pointa, ktorá uniká postmoderným ľuďom kričiacim: „Ale ako môžete *vedieť*, či sú vaše názory pravdivé?“ Keď robíte experiment, v skutočnosti *vychádzate* von zo svojej hlavy. Zúčastňujete sa na zložitej interakcii, ktorej výsledok je kauzálne určený vecou, o ktorej rozmýšľate, nie iba vašimi názormi o nej. Raz som definoval „skutočnosť“ takto:

> Dokonca aj keď mám jednoduchú hypotézu, silno podporovanú všetkými indíciami, ktoré poznám, niekedy som aj tak prekvapený. Potrebujem teda odlišné názvy na veci, ktoré určujú moje predpovede a veci, ktoré určujú moje experimentálne výsledky. To prvé nazývam „názory“ a to druhé „skutočnosť“.

Interpretácia vášho pokusu stále závisí od vašich pôvodných názorov. Nebudem zatiaľ hovoriť o tom, Odkiaľ Pochádzajú Pôvodné Názory, pretože to nie je témou tejto kapitoly. Pointa je, že pravda odkazuje na *ideálne* porovnanie medzi názorom a skutočnosťou. Vďaka tomu, že rozumieme, že planéty sú niečo iné ako názory o planétach, dokážeme zostaviť experiment na otestovanie, či názor „zornička a večernica je tá istá planéta“ je *pravdivý*. Tento pokus bude zahŕňať ďalekohľady, nie iba introspekciu, pretože rozumieme, že „pravda“ zahŕňa porovnávanie vnútorného názoru s vonkajším faktom; takže použijeme prístroj, ďalekohľad, ktorého vnímané správanie podľa nás závisí od vonkajšieho faktu planéty.

Dôvera, že nám ďalekohľad pomáha vyhodnotiť „pravdu“ tvrdenia „zornička = večernica“, sa zakladá na našich pôvodných názoroch o tom, ako ďalekohľad interaguje s planétou. Opäť, nebudem sa

tomu venovať v tejto kapitole, okrem citovania jedného z mojich obľúbených výrokov Raymonda Smullyana: „Ak sofistikovanejší čitateľ namieta voči tomuto výroku na základe toho, že je to púha tautológia, nech prosím tomuto výroku uzná aspoň to, že nie je nekonzistentný.“ Podobne, nevidím používanie ďalekohľadu ako kruhovú logiku, ale ako reflektívnu koherenciu; pre každý systematický spôsob získavania pravdy by malo existovať rozumné vysvetlenie, prečo to funguje.

Aktuálna otázka je, čo to *znamená*, že „sneh je biely“ je *pravda*, keď v skutočnosti existujú iba kvarky.

Existuje istý vzor neurónových spojení, ktoré vytvárajú váš názor na „sneh“ a „bielosť“ - tomuto veríme, aj keď si nevieme a nedokážeme konkrétne predstaviť tieto skutočné neurónové spojenia. Ktoré samotné sú vytvorené ako vzor ešte menej známych kvarkov. Tam vo svete existujú molekuly vody, ktorých teplota je dostatočne nízka, aby sa usporiadali do pravidelne sa opakujúcich vzorov; vôbec sa nepodobajú na motanicu neurónov. V akom zmysle, keď porovnávame jeden (stále sa meniaci) vzor kvarkov s druhým, je názor „sneh je biely“ *pravdivý*?

Je zrejmé, že ani ja ani nikto iný nedokáže ponúknuť Ideálnu Funkciu Porovnania Kvarkov, ktorá by dostala popis názoru stelesneného v neurónoch (vrátane okolitého mozgu) na úrovni kvarkov a popis snehovej vločky (a príslušných zákonov optiky) na úrovni kvarkov, a vypísala by „pravda“ alebo „nepravda“ na výrok „sneh je biely“. A kto hovorí, že základná úroveň *naozaj* sú polia častíc?

Na druhej strane, vyhodiť všetky naše názory preto, lebo nie sú napísané ako gigantická nezvládnuteľná špecifikácia o neviditeľných kvarkoch... nevyzerá ako veľmi prezieravý nápad. Ani ako najlepší spôsob ako optimalizovať naše ciele.

Zdá sa mi, že slová ako „sneh“ a „biely“ možno brať ako istý druh dlžného úpisu – nie ako *známu* špecifikáciu, ktoré presne zostavy fyzických kvarkov sa počítajú za „sneh“, ale predsa len, že existujú veci, ktoré nazývate sneh, a existujú veci, ktoré nenazývate sneh, a aj keby ste sa v pár veciach pomýlili (napríklad umelý sneh), Ideálny Vševediaci Interpreter Vedy by videl hustý zhluk v strede a prekreslil by hranicu, aby mala jednoduchšiu definíciu.

V jednoúrovňovom vesmíre, ktorého spodnú úroveň nepoznáme alebo je nejasná alebo skrátka príliš veľká na to, aby sme o nej hovorili, sa o pojmoch viacúrovňovej mysle dá povedať, že predstavujú istý druh dlžného úpisu – nevieme, *čomu* tam vonku zodpovedajú. Ale zdá sa nám, že dokážeme odlíšiť pozitíve prípady od negatívnych predvídateľne produktívnym spôsobom, a tak si myslíme – možno v plne všeobecnom zmysle – že existuje *nejaký* rozdiel v kvarkoch, *nejaký* rozdiel v konfiguráciách na základnej úrovni, ktorý vysvetľuje rozdiely, ktoré sa dostávajú do našich zmyslov, a v konečnom dôsledku spôsobujú, že povieme „sneh“ alebo „nie sneh“.

Vidím túto bielu hmotu a je taká istá v niekoľkých prípadoch, takže predpokladám stabilnú skrytú príčinu v prostredí – nazývam ju „sneh“; „sneh“ je potom dlžný úpis odkazujúci na predpokladanú jednoduchú hranicu, ktorú možno nakresliť okolo nevidených príčin môjho vnemu.

Myšlienkový pokus Hilary Putnamovej o „dvojčati Zeme“, kde voda nie H2O alebo nejaká čudná iná látka označená XYZ, ktorá sa inak správa celkom ako voda, a následná filozofická debata pomáha zdôrazniť túto tému. „Sneh“ nemá nám známu logickú definíciu – je to skôr empiricky určený ukazovateľ na logickú definíciu. To je pravda aj keď veríte, že sneh sú ľadové kryštály, čo sú molekuly vody usporiadané pri nízkej teplote. Molekuly vody sa skladajú z kvarkov. Čo ak sa ukáže, že sa kvarky skladajú z niečoho ďalšieho? Čo *je* potom snehová vločka? Neviete – ale stále je to snehová vločka, nie požiarny hydrant.

A samozrejme, samotné tieto odseky, ktoré som práve napísal, sú podobne vysoko nad úrovňou kvarkov. „Vnímať bielu vec, vizuálne ju zaradiť, a myslieť si ‚sneh‘ alebo ‚nie sneh‘“ - aj toto je rozprávanie veľmi vysoko nad úrovňou kvarkov. Takže moje meta-názory sú tiež dlžnými úpismi pre veci, o ktorých by Ideálny Vševediaci Interpreter Vedy mohol vedieť, ktoré zostavy kvarkov (alebo hocičoho) tvoriacich môj mozog zodpovedajú „názoru, že ‚sneh je biely‘“.

Ale potom všetko, čo zo skutočnosti vieme uchopiť, sa skladá z dlžných úpisov tohto druhu. Takže namiesto označovania za kruhové to radšej označujem za vnútorne konzistentné.

Toto môže trochu znervózňovať – držať si vratké epistemické vyvýšené miesto na úrovni aj názorov o predmetoch, aj reflexie, vysoko nad obrovskou základnou skutočnosťou za tým všetkým, a dúfať, že odtiaľ nespadneme.

Po reflexii je však ťažko navrhnúť, ako inak by to mohlo byť.

Na konci dňa teda výrok „skutočnosť neobsahuje ruky ako základné, pridané, samostatné kauzálne entity, navyše ku kvarkom“ nie je to isté ako výrok „ruky neexistujú“ alebo „nemám ruky“. Neexistujú žiadne *fundamentálne* ruky; ruky sa skladajú z prstov a dlaní, ktoré sa zase skladajú zo svalov a kostí, a tak ďalej až po polia základných častíc, ktoré podľa toho, čo teraz vieme, sú základnými kauzálnymi entitami.

To nie je to isté ako povedať: „‚ruky‘ neexistujú“. Nie to to isté ako povedať: „slovo ‚ruky‘ je dlžný úpis, ktorý nikdy nebude vyplatený, pretože neexistuje zodpovedajúci empirický zhluk“; alebo „zmenka ‚ruky‘ sa nikdy nevyplatí, pretože je logicky nemožné uviesť do súladu jej domnelé vlastnosti“; alebo „veta ‚ľudia majú ruky‘ odkazuje na zmysluplný stav vecí, ale skutočnosť nie je v tomto stave“.

Iba: Tam, kde my vidíme „ruky“, existujú *v* skutočnosti vzory, a tieto vzory majú čosi spoločné, ale nie sú fundamentálne.

Keby som *naozaj* nemal ruky – keby sa skutočnosť náhle prepla do stavu, ktorý by sme opísali ako: „Eliezer nemá ruky“ - skutočnosť by krátko na to zodpovedala stavu, ktorý by sme popísali ako: „Eliezer vrieska, zatiaľ čo mu krv strieka z kýpťov na ramenách“.

A to je *pravda*, aj keď predchádzajúci odsek nešpecifikoval žiadne polohy kvarkov.

Predchádzajúca veta je takisto meta-pravdivá.

Mapa je viacúrovňová, skutočnosť má jednu úroveň. To neznamená, že tie vyššie úrovne „neexistujú“ ako keď vo svoje garáži hľadáte draka a nenájdete tam nič, alebo ako keď vidíte fatamorgánu v púšti a vytvoríte si očakávanie pitnej vody, keď sa tam nedá nič piť. Vyššie úrovne vašej mapy nie sú *nepravdivé*, bez referenta; majú referenty *v* tej jednej úrovni fyziky. Nie je to tak, že by lietadlo nemalo krídla – potom by spadlo z oblohy. „Krídla lietadla“ existujú *explicitne* v inžinierovom viacúrovňovom modeli lietadla, a krídla lietadla existujú *implicitne* v kvantovej fyzike skutočného lietadla. Existovať implicitne nie je to isté čo neexistovať. Presný popis tejto implicitnosti nepoznáme – nie je explicitne reprezentovaný na našej mape. Ale to našej mape nebráni, aby fungovala, dokonca ani aby bola *pravdivá*.

Hoci je trochu znervózňujúce zamýšľať sa nad tým, že každý jeden pojem a názor vo vašom mozgu, vrátane týchto meta-pojmov o tom, ako váš mozog funguje a prečo si dokážete vytvárať správne názory, sa týči rády a rády nad skutočnosťou...

*→

221. Zombie! Zombie?

Vaša „zombia“ vo filozofickom použití tohto pojmu, je údajná bytosť, ktorá sa správa v *každom* ohľade rovnako ako vy – rovnaké správanie, rovnaká reč, rovnaký mozog; každý atóm a kvark v *presne* rovnakej polohe, pohybujúci sa podľa rovnakých kauzálnych zákonov pohybu – *akurát* že vaša zombia nemá vedomie.

Ďalej sa tvrdí, že ak sú zombie „možné“ (čo je pojem, o ktorý sa stále bojuje), potom zo samotného nášho poznania tejto „možnosti“ môžeme a priori odvodiť, že vedomie je mimo fyziky, v zmysle, ktoré popíšem nižšie; štandardný pojem pre tento postoj je „epifenomenalizmus“.

(Pre tých, ktorí nie sú oboznámení so zombiami, zdôrazňujem, že *toto nie je slamený panák*. Pozrite si napríklad heslo SEP pre Zombie. „Možnosť“ zombií prijíma podstatná časť, možno väčšina akademických filozofov vedomia.)

Kdesi som čítal: „Nie si ten, kto hovorí tvoje myšlienky – si ten, kto tvoje myšlienky *počuje*.“ V hebrejčine slovo pre najvyššiu dušu, ktorú Boh vdýchol do Adama, je N'Shama - „počujúci“.

→ http://lesswrong.com/lw/p6/reductive_reference/

Ak si predstavíte „vedomie“ ako čisto pasívne počúvanie, potom je predstava zombií na začiatku veľmi ľahká. Je to niekto, kto nemá N'Shama, počúvajúceho.

(Upozornenie: Nasleduje *veľmi* dlhý článok so 6600 slovami zahŕňajúci Davida Chalmersa. Môžete to brať ako môj ukážkový protipríklad na článok Richarda Chappella Hádanie sa s Eliezerom, časť II, v ktorej ma Richard obviňuje, že sa nevenujem zložitým argumentom skutočných filozofov.)

Keď otvoríte chladničku a zistíte, že sa minul pomarančový džús, pomyslíte si: „Sakra, minul sa pomarančový džús.“ Zvuk týchto slov sa pravdepodobne reprezentuje niekde vo vašej zvukovej kôre, akoby ste to počuli povedať niekoho iného. (Prečo si to myslím? Pretože rodení čínsky hovoriaci si dokážu zapamätať dlhšie postupnosti číslic ako anglicky hovoriaci. Čínske číslice sú všetky jednoslabičné, a tak si čínsky hovoriaci dokážu zapamätať okolo desať číslic, v porovnaní so známym „sedem plusmínus dva“ pre anglicky hovoriacich. Zdá sa, že existuje slučka opakovania zvukov pre seba, hranica veľkosti v pracovnej pamäti sluchovej kôry, ktorá sa naozaj zakladá na hláskach.)

Predpokladajme, že horeuvedené je pravda; ako tvrdenie by to iste pre zástancu zombií nepredstavovalo problém. Aj keby ľudia takíto neboli, zdá sa dosť ľahké predstaviť si takto zostavenú UI (a predstaviteľnosť je podstatou celej debaty o zombiách). Nie je iba principiálne predstaviteľné, ale pomerne možné, že v najbližších desaťročiach chirurgovia niekomu rozložia sieť neurónových senzorov po sluchovej kôre a prečítajú si jeho vnútorný príbeh. (Výskumníci už zaznamenávali bočné genikulátne teleso mačky a rekonštruovali rozoznateľné zrakové vstupy.)

Takže vaša zombia, keďže je s vami fyzikálne totožná do posledného atómu, otvorí chladničku a vytvorí sluchové kôrové vzory pre hlásky: „Sakra, minul sa pomarančový džús.“ V tomto bode by epifenomenalisti ochotne súhlasili.

Lenže, povie epifenomenalista, u zombie nie je nikto vnútri, kto by to *počul*; chýba vnútorný poslucháč. Vnútorný príbeh sa rozpráva, ale nepočúva. Vy nie ste ten, kto hovorí vaše myšlienky, ale ten, kto ich počuje.

Zdá sa omnoho priamočiarejšie (povedali by) vytvoriť UI, ktorá by vypisovala nejaký druh vnútorného príbehu, než ukázať, že ho počúva vnútorný poslucháč.

Argument zombií je, že ak je svet zombií *možný* – nie nevyhnutne fyzikálne možný v našom vesmíre, iba „teoreticky možný“ alebo „predstaviteľný“, alebo niečo v tom duchu – potom vedomie musí byť mimofyzikálne, niečo viac a iné než púhe atómy. Prečo? Pretože aj keby ste akosi poznali polohy všetkých atómov vo vesmíre, stále by vám bolo treba povedať, ako samostatný a oddelený fakt, že ľudia majú vedomie – že majú vnútorných poslucháčov – že nie sme vo svete zombií, ako sa zdá *možné*.

Zombi-izmus nie je to isté ako dualizmus. Descartes si myslel, že existuje telesná hmota a celkom iný druh myšlienkovej hmoty, ale Descartes si myslel aj, že myšlienková hmota je *kauzálne aktívny* princíp, interagujúci s telesnou hmotou, ovládajúci našu reč a správanie. Odobranie myšlienkovej hmoty z človeka by ponechalo *tradičnú* zombiu, potácajúcu sa a vzdychajúcu.

A hoci je hebrejské slovo pre najvnútornejšiu dušu N'Shama, to-čo-počuje, nespomínam si, že by som počul nejakého rabína argumentovať za možnosť zombií. Väčšina rabínov by pravdepodobne bola zhrozená z predstavy, že božská časť, ktorú Boh vdýchol do Adama, *v skutočnosti nič nerobí*.

Technický pojem pre vieru, že vedomie existuje, ale nemá žiaden účinok na fyzikálny svet, je *epifenomenalizmus*.

Argument zombiami má aj ďalšie prvky (ktorým sa budem venovať nižšie), ale myslím si, že je to intuícia pasívneho poslucháča, ktorá ako prvá zvádza ľudí na zombi-izmus. Konkrétne je to to, čo zvádza na zombi-izmus laické obecenstvo. Ústredný pojem je ľahký a dostupný: Svetlá sú zapnuté, ale nikto nie je doma.

Filozofi sa odvolávajú na intuíciu pasívneho poslucháča, keď hovoria: „Samozrejme, že si svet zombií možno predstaviť; presne viete, ako by to vyzeralo.“

Jedna z veľkých bitiek vo Vojnách Zombií je o tom, čo presne sa myslí slovami, že zombie sú „možné“. Raní zombi-istickí filozofi (1970-te roky) si iba mysleli, že je zrejmé, že zombie sú „možné“, a neobťažovali sa definovať, aký druh možnosti tým myslia.

Pretože som čítal o matematickej logike, okamžite mi napadá logická možnosť. Ak máte zbierku výrokov ako (A -> B), (B -> C), (C -> ~A), potom je zložený názor *logicky možný* vtedy, keď má *model* – čo sa v horeuvedenom prípade redukuje na nájdenie hodnôt, ktoré priradíme A, B, C, aby všetky výroky (A -> B), (B -> C) a (C -> ~A) boli pravdivé. V tomto prípade funguje A = B = C = 0, ale aj A = 0, B = C = 1 alebo A = B = 0, C = 1.

Niečo bude *vyzerať* možné – bude „pojmovo možné“ alebo „predstaviteľné“ - ak sa viete zamyslieť nad skupinou výrokov bez toho, že by ste *videli* rozpor. Ale vo všeobecnosti je veľmi ťažký problém vidieť rozpory *alebo* nájsť celkom konkrétny model! Ak sa obmedzíte na jednoduché Boolovské výroky tvaru ((A alebo B alebo C) a (B alebo ~C alebo D) a (D alebo ~A alebo ~C) …), konjunkcie disjunkcií troch premenných, toto je veľmi známy problém zvaný 3-SAT, čo je jeden z prvých problémoch, o ktorých sa dokázalo, že sú NP-úplné.

Takže to, že vy na prvý pohľad nevidíte rozpor vo Svete Zombií, neznamená, že tam rozpor nie je. Je to ako nevidieť na prvý pohľad rozpor v Riemannovej Hypotéze. Od pojmovej možnosti („nevidím tam problém“) po *logickú možnosť* v plnom technickom zmysle slova je veľmi veľký skok. Je jednoduché urobiť z toho NP-úplný skok, a pomocou teórií prvého rádu môžete urobiť aj počítanie konečných otázok neobmedzene zložité. A potrebujeme *logickú možnosť* Sveta Zombií, nie pojmovú možnosť, na predpoklad, že by logicky vševedúca myseľ mohla poznať polohy všetkých atómov vo vesmíre a predsa by jej bolo treba povedať *dodatočný* ešte nezahrnutý fakt, že máme vnútorných poslucháčov.

Iba to, že *zatiaľ* nevidíte rozpor, nie je zárukou, že neuvidíte rozpor o ďalších 30 sekúnd. „Všetky nepárne čísla sú prvočísla. Dôkaz: 3 je prvočíslo, 5 je prvočíslo, 7 je prvočíslo...“

Zamyslime sa teda nad argumentom zombií *o trochu dlhšie*: Vieme si predstaviť protipríklad k tvrdeniu: „Vedomie nemá žiaden kauzálny dopad na svet zistiteľný treťou stranou?“

Ak zavriete oči a sústredíte sa na svoje vnútorné vedomie, začnú sa vo vašom vnútornom príbehu tvoriť myšlienky ako „Uvedomujem si“ a „Moje vedomie je oddelené od mojich myšlienok“ a „Ja nie som ten, kto hovorí moje myšlienky, ale ten, kto ich počuje“ a „Môj prúd vedomia nie je moje vedomie“ a „Vyzerá to, že existuje časť mňa, ktorej odstránenie si viem predstaviť bezo zmeny môjho vonkajšieho správania“.

Môžete tieto vety dokonca povedať nahlas, ako rozjímate. V princípe by niekto so super-fMRI asi mohol prečítať tieto hlásky priamo z vašej sluchovej kôry; ale povedať ich nahlas odstraňuje všetky pochybnosti o tom, či ste stúpili do sveta testovateľnosti a fyzikálnych dôsledkov.

Toto iste vyzerá ako vášho vnútorného poslucháča *pristihla pri akte počúvania* tá časť vás, ktorá zapisuje vnútorný príbeh a trepoce vaším jazykom.

Predstavte si, že vás navštívi tajomná mimozemská rasa a nechá vám ako dar tajomnú čiernu krabicu. Pokúšate sa do krabice pichať a štuchať, ale (pokiaľ viete) nikdy sa vám nepodarilo vyvolať reakciu. Nedokážete dosiahnuť, aby čierna krabica produkovala zlaté mince alebo odpovedala na otázky. Preto usúdite, že táto čierna krabica je kauzálne neaktívna: „Pre každé X platí, že čierna krabica nerobí X.“ Čierna krabica je účinok, ale nie príčina; epifenomenálna; nemá kauzálnu potenciu. Vo svojej mysli otestujete túto všeobecnú hypotézu, aby ste videli, či je pravdivá v niekoľkých pokusných prípadoch, a zdá sa pravdivá. - „Premieňa čierna krabica olovo na zlato? Nie. Zohrieva čierna krabica vodu? Nie.“

Lenže vy túto čiernu krabicu *vidíte*; pohlcuje svetlo, a zaťažuje vašu ruku. Aj toto je súčasťou tanca kauzality. Keby bola čierna krabica *celkom* mimo kauzálneho vesmíru, nevideli by ste ju; nemali by ste ako vedieť, že existuje; nemohli by ste povedať: „Ďakujem za čiernu krabicu.“ Na tento protipríklad ste *nepomysleli*, keď ste formulovali všeobecné pravidlo: „Pre každé X: Čierna krabica nerobí X.“ Ale bolo to stále tam.

(V skutočnosti vám mimozemšťania nechali *inú* čiernu krabicu, tentokrát *celkom* epifenomenálnu, a nemali ste najmenšie tušenie o tom, že je vo vašej obývačke. Toto bol ich vtip.)

Ak môžete zavrieť oči a cítiť, že vnímate – ak si dokážete uvedomovať, že si uvedomujete, a pomyslieť si: „Uvedomujem si, že si uvedomujem“ - a povedať nahlas: „Uvedomujem si, že si uvedomujem“ - potom vaše vedomie nie je bez účinku na váš vnútorný príbeh, alebo na pohyb vašich

pier. Môžete vidieť ako vidíte, a vaše vnútorné rozprávanie toto odráža, a aj vaše pery, ak sa to rozhodnete povedať nahlas.

Horeuvedený argument som nevidel napísaný týmto konkrétnym spôsobom - „poslucháč prichytený pri akte počúvania“ - ale možno to už niekto predo mnou povedal.

Je však štandardným bodom – ktorý zombi-istickí filozofi akceptujú! - že filozofii vo Svete Zombií, keďže sú atóm za atómom zhodní s našimi vlastnými filozofmi, píšu totožné články o filozofii vedomia.

V tomto bode svet zombií prestáva byť intuitívnym dôsledkom predstavy pasívneho poslucháča.

Filozofi píšuci články o vedomí, to *vyzerá* ako prinajmenšom jeden účinok vedomia na svet. Môžete argumentovať chytré dôvody, prečo to tak nie je, ale musíte byť chytrý.

Mohli by ste intuitívne predpokladal, že keby vaše vnútorné vedomie zmizlo, svet by sa zmenil v tom, že by vaše vnútorné rozprávanie už nehovorilo veci ako: „Je vo mne tajomný poslucháč“, pretože ten tajomný poslucháč by bol preč. Zvyčajne hned *po tom*, ako zameriate svoje vedomie na svoje vedomie, vaše vnútorné rozprávanie povie: „Uvedomujem si, že si uvedomujem“, čo naznačuje, že keby sa nebola stala tá prvá udalosť, nebola by sa stala ani tá druhá. Môžete chytro argumentovať, prečo to tak nie je, ale musíte byť chytrý.

Môžete vytvoriť výrokový názor, že „Vedomie je bez účinku“, a *nevidieť* na prvý pohľad žiaden rozpor, ak si neuvedomujete, že hovoriť o vedomí je účinok vedomia. Ale keď už vidíte spojenie medzi všeobecným pravidlom, že vedomie nemá žiaden účinok, a konkrétnym dôsledkom, že vedomie nemá žiaden účinok na to, ako filozofi píšu články o vedomí, zombi-izmus prestane byť intuitívny a začne si vyžadovať tvrdenie zvláštnych vecí.

Jednou zvláštnou vecou, ktorú by ste mohli tvrdiť, je existencia vládcu zombií, boha vo svete zombií, ktorý tajne ovláda zombie filozofov a núti ich hovoriť a písať o vedomí.

Vládca zombií nevyzerá byť nemožný. Ľudia často neznejú celkom koherentne, keď hovoria o vedomí. Možno by nebolo taká ťažké napodobniť ich diskusie, povedzme na úrovni amatérskeho človeka hovoriaceho v krčme. Možno by ste mohli vziať ako korpus tisíc ľudských amatérov, ktorí skúšajú diskutovať o vedomí; nakŕmiť tým nevedomú ale sofistikovanú UI, lepšiu než dnešné modely, ale nie sebamodifikujúcu; a získať naspäť pojednanie o „vedomí“, ktoré bude znieť rovnako zmysluple ako väčšina ľudí, čo znamená, nie veľmi.

Ale všetka táto reč o „vedomí“ by nebola spontánna. Nebola by vytvorená *v rámci* UI. Bola by to nahraná napodobenina hovorenia niekoho iného. Toto je iba holodeck, kde ústredná UI píše prejavy nehráčskych postáv. Toto *nie je* to, o čom je svet zombií.

Podľa predpokladu je svet zombií atóm za atómom totožný s naším svetom, akurát že jeho obyvatelia nemajú vedomie. Navyše, atómy vo svete zombií sa pohybujú podľa rovnakých fyzikálnych zákonov ako v našom svete. Ak existujú „premosťovacie zákony“, ktoré riadia, *ktoré zostavy atómov vyvolávajú vedomie*, tieto premosťovacie zákony tam nie sú. Ale podľa hypotézy tento rozdiel nie je experimentálne zistiteľný. Keď príde na to, či nejaký kvark urobí cik alebo cak alebo či zapôsobí silou na susedné kvarky – čokoľvek experimentálne merateľné – platia tie isté zákony fyziky.

Vo svete zombií nie je *miesto* pre vládcu zombií, pretože vládca zombií musí ovládať pery zombií, a takáto kontrola je v princípe experimentálne zistiteľná. Vládca zombií pohybuje pery, preto má pozorovateľné dôsledky. Existoval by bod, kde elektrón urobí cik namiesto cak, pretože tak povedal vládca zombií. (Pokiaľ teda vládca zombií nie je *súčasťou* toho sveta, ako vzor kvarkov – ale potom by svet zombií nebol atóm po atóme totožný s naším vlastným, pokiaľ si nemyslíte, že aj *tento* svet obsahuje vládcu zombií.)

Keď filozof v našom svete napíše: „Myslím si, že svet zombií je možný,“ jeho prsty postupne stláčajú klávesy Z-O-M-B-I-Í. Existuje reťaz kazuality, ktorú možno vystopovať späť z týchto stlačení kláves: stiahnutia svalov, signály nervov, príkazy poslané dole miechou, z pohybovej kôry – a potom do menej známych oblastí mozgu, kde filozofove vnútorné rozprávanie prvýkrát začalo hovoriť o „vedomí“.

A filozofove zombícke dvojča stláča rovnaké klávesy, *z rovnakého dôvodu*, kauzálne povedané. V tejto reťazi vysvetlení neexistuje žiadna príčina, prečo filozof píše tak, ako píše, ktorá by nebola prítomná aj u jeho zombíckeho dvojčaťa. Aj zombícke dvojča má vnútorné rozprávanie o „vedomí“, ktoré by

super-fMRI mohlo prečítať z jeho sluchovej kôry. A ľubovoľné iné myšlienky alebo ľubovoľné iné príčiny, ktoré viedli k vnútornému rozprávaniu, sú presne rovnaké v našom vesmíre aj vo svete zombií.

Nemôžete teda povedať, že filozof píše o vedomí *pretože* má vedomie, zatiaľ čo jeho zombícke dvojča píše o vedomí kvôli vládcovi zombií alebo chatujúcej UI. Keď vystopujete späť reťaz kauzality za klávesnicou, k vnútornému rozprávaniu opakovanom v sluchovej kôre, k príčine tohto rozprávania, musíte nájsť *rovnaké* fyzikálne vysvetlenie v našom svete ako vo svete zombií.

Ako píše najobávanejší zástanca zombi-izmu, David Chalmers:[203]

> Pomyslime na moje zombícke dvojča vo vedľajšom vesmíre. Stále hovorí o vedomej skúsenosti – v skutočnosti tým vyzerá byť posadnutý. Trávi absurdné množstvá času zhrbený nad počítačom, píše kapitolu za kapitolou o tajomstvách vedomia. Často komentuje o pôžitku, ktorý má z istých zmyslových kválií, vyjadruje osobitnú náklonnosť k výraznej zelenej a fialovej. Často sa dostane do sporu so zombiami materialistami, argumentujúc, že ich postoje nezodpovedajú skutočnosti vedomej skúsenosti.
>
> A predsa nemá vôbec žiadnu vedomú skúsenosť! V jeho vesmíre majú materialisti pravdu a on sa mýli. Väčšina jeho tvrdení o vedomej skúsenosti je celkom nesprávna. Ale iste existuje fyzikálne alebo funkcionálne vysvetlenie, prečo robí tieto závery. Napokon, jeho svet je plne riadený zákonmi, a nie sú tam žiadne zázračné udalosti, takže musí existovať nejaké vysvetlenie jeho záverov.
>
> ...Ľubovoľné vysvetlenie správania môjho dvojčaťa sa bude rovnako počítať ako vysvetlenie môjho správania, pretože procesy v jeho tele presne zodpovedajú tým v mojom tele. Vysvetlenie jeho záverov očividne nezávisí na existencii vedomia, lebo v jeho svete neexistuje žiadne vedomie. Z toho vyplýva, že vysvetlenie mojich záverov tiež nezávisí na existencii vedomia.

Chalmers tu neargumentuje *proti* zombiám; toto sú naozaj jeho názory!

> Táto paradoxná situácia je zároveň pôvabná i znepokojujúca. Nie je očividne osudná pre nereduktívny postoj, ale je tu aspoň niečo, s čím sa musíme popasovať...

Pri všetkej vážnosti by som toto nominoval za najväčšie postavenie sa guľke do cesty v dejinách času. A to je dvojsečný kompliment pre Davida Chalmersa: Menší smrteľník by si jednoducho nevšimol dôsledky, alebo by im odmietol čeliť, alebo by si racionalizoval dôvod, prečo to tak nie je.

Prečo by sa niekto postavil do cesty takejto veľkej guľke? Prečo by niekto predpokladal nevedomé zombie, ktoré píšu články o vedomí z *celkom rovnakého dôvodu* ako náš vlastný skutočne vedomý filozof?

Nie kvôli ten prvej intuácii, o ktorej som písal, intuícii pasívneho poslucháča. Táto intuícia môže povedať, že zombie môžu šoférovať autá alebo počítať matematiku alebo sa dokonca zamilovať, ale nehovorí, že zombie môžu písať filozofické články o svojich pasívnych posluchačoch.

Argument zombií nespočíva *výhradne* na intuícii pasívneho poslucháča. Keby toto bolo všetko, na čom argument zombií stojí, už by bolo dávno po ňom, myslím si. Intuícia, že môžeme „poslucháča“ odstrániť bez dôsledkov, by pominula pri uvedomení si, že vaše vnútorné rozprávanie pravidelne *vyzerá*, že zachytáva poslucháča v akte počúvania.

Nie, aby sa niekto postavil do cesty *tejto* guľke, musí prísť s celkom inou intuíciou – s intuíciou, že bez ohľadu na to, koľko atómov dáte dokopy, bez ohľadu na to, koľko hmoty a elektrického náboja navzájom interaguje, nikdy *nevyhnutne* nevytvoria subjektívny vnem tajomnej červenosti červenej. Môže to byť fakt o našom fyzikálnom vesmíre (hovorí Chalmers), že dať také a také atómy do takej a takej polohy *vyvoláva* pocit červenosti; ale ak je to tak, nie je to *nevyhnutný* fakt, je to niečo, čo si vyžduje vysvetlenie mimo pohybu atómov.

Ale ak sa zamyslíte nad touto druhou intuíciou osamote, bez intuície pasívneho poslucháča, je ťažké vidieť, prečo z nej vyplýva zombi-izmus. Možno len existuje *iný druh hmoty*, okrem atómov, ktorý *nie je*

203 Chalmers, *The Conscious Mind.*

kauzálne pasívny – duša, ktorá naozaj niečo *robí*, duša, ktorá hrá skutočnú kauzálnu rolu v tom, prečo píšeme o „tajomnej červenosti červenej“. Odoberte túto dušu a... nuž, pokiaľ neupadnete do kómy, určite nebudete písať ďalšie články o vedomí!

Toto je postoj, ktorý zaujal Descartes a mnohí ďalší dávni myslitelia: Duša je iného druhu, ale *interaguje* s telom. Descarov postoj sa technicky označuje ako *dualizmus hmoty* – existuje predmetová hmota, a myšlienková hmota, ktoré nie je ako atómy; ale má kauzálnu schopnosť, interaguje, a zanecháva na našom svete viditeľnú stopu.

Zombi-isti sú *dualisti vlastností* – neveria v oddelenú dušu; veria v to, že hmota v našom vesmíre má *ďalšie vlastnosti* okrem fyzikálnych.

„Okrem fyzikálnych?“ To čo znamená? Znamená to, že tie vlastnosti navyše sú tam, ale neovplyvňujú pohyb atómov, na rozdiel od vlastností ako elektrický náboj alebo hmotnosť. Tieto vlastnosti navyše sa nedajú pokusne zachytiť *treťou stranou*; vy viete, že máte vedomie, *zvnútra* vašich vlastností navyše, ale žiaden vedec to nikdy nemôže priamo zistiť zvonka.

Takže tie dodatočné vlastnosti sú tam, ale nie sú kauzálne aktívne. Tie vlastnosti navyše nepohybujú atómy, a preto ich nemôže zaznamenať tretia strana.

A preto si dokážeme (údajne) predstaviť vesmír presne ako ten náš, kde sú všetky atómy na rovnakých miestach, ale tieto vlastnosti navyše chýbajú, takže všetko sa pohybuje rovnako ako predtým, ale nikto nemá vedomie.

Svet zombií nemusí byť *fyzikálne* možný, hovoria zombi-isti – pretože je faktom, že všetky hmota v našom vesmíre má tieto vlastnosti navyše, alebo poskúcha premosťovacie zákony, ktoré vyvolávajú vedomie – ale svet zombií je *logicky* možný: tie premosťovacie tákony by mohli byť iné.

Ale akonáhle si uvedomíte, že predstaviteľnosť nie je to isté ako logická možnosť, a že svet zombií nie je ani celkom intuitívny, prečo hovoríme, že svet zombií je logicky možný?

Prečo, ach prečo, hovoríme, že tieto vlastnosti navyše sú epifenomenálne a nezistiteľné?

Môžeme túto dilemu položiť veľmi ostro: Chalmers verí, že *existuje* niečo, čo sa volá vedomie, a že toto vedomie stelesňuje pravú a neopísateľnú podstatu tajomnej červenosti červenej. Môže to byť vlastnosť mimo hmotnosti a náboja, ale je *tam*, a *je* to vedomie. Po tom, čo povedal horeuvedené, Chalmers ďalej upresňuje, že táto skutočná vec vedomia je epifenomenálna, bez kauzálnej schopnosti – ale *prečo to hovorí*?

Prečo hovorí, že by ste mohli odobrať túto pravú vec vedomia a nechať všetky atómy na rovnakom mieste, a robili by rovnaké veci? Ak je to pravda, potom potrebujeme nejaké *osobitné* vysvetlenie, prečo Chalmers hovorí o „tajomnej červenosti červenej“. To znamená, že existuje aj tajomná červenosť červenej, ktorá je mimofyzikálna, aj *celkom nezávislý* dôvod, *v rámci* fyziky, prečo Chalmers hovorí o „tajomnej červenosti červenej“.

Chalmer priznáva, že tieto dve veci vyzerajú akoby navzájom súviseli, ale v skutočnosti, načo potrebujeme obe? Prečo si jednoducho nevybrať jednu alebo druhú?

Keď ste raz predpokladali, že existuje tajomná červenosť červenej, prečo nepovedať, že interaguje s vaším vnútorným rozprávaním a spôsobuje, že hovoríte o „tajomnej červenosti červenej“?

Nepoužíva tu Descartes jednoduchší prístup? *Striktne* jednoduchší prístup?

Prečo predpokladať nehmotnú dušu, *a potom* predpokladať, že táto duša nemá žiaden účinok na fyzikálny svet, *a potom* predpokladať tajomný neznámy *hmotný* proces, ktorý spôsobuje, že váš vnútorný rozhovor hovorí o vedomej skúsenosti?

Prečo nepredpokladať pravú vec vedomia, ktorá sa nedá zložiť zo žiadneho množstva púhych mechanických atómov, *a potom*, keď už sme zašli takto ďaleko, nenechať túto pravú vec vedomia mať kauzálne účinky, napríklad spôsobovať, že filozofi hovoria o vedomí?

Nepodporujem tu Descartov pohľad. Ale prinajmenšom dokážem pochopiť, z čoho Descartes vychádza. Vedomie vyzerá byť tajomné, preto predpokladáte tajomnú látku vedomia. Dobre.

Ale zombi-isti teraz predpokladajú, že táto tajomná vec *nerobí nič*, takže potrebujete ešte *celkom nové* vysvetlenie, prečo *hovoríte*, že máte vedomie.

Toto nie je vitalizmus. Toto je niečo také čudné, že vitalisti by to vypľuli. „Keď oheň horí, uvoľňuje flogiston. *Ale* flogiston nemá žiaden experimentálne zistiteľné účinok na náš vesmír, takže musíme hľadať *nezávislé* vysvetlenie, prečo oheň dokáže roztopiť sneh.“ *Čože?*

Majú dualisti vlastností dojem, že keď budú predpokladať novú *aktívnu* silu, niečo, čo má kauzálny dopad na pozorovateľné veci, že budú už príliš vystrkovať hlavu?

Ja by som povedal, že ak predpokladáte tajomnú, nezávislú, dodatočnú, vnútorne myšlienkovú vlastnosť vedomia, mimo polôh a rýchlostí, v tomto bode ste *už* vystrčili hlavu tak ďaleko, ako sa len dá. Aby ste predpokladali túto vec vedomia, a potom navyše predpokladali, že *nerobí nič* – pre lásku milých mačiatok, *prečo*?

Dokonca tam nie je ani jasný kariérny motív. „Dobrý deň, som filozof vedomia. Moja téma je tá najdôležitejšia vec v celom vesmíre, nemal by som dostať veľa grantov? To je pekné, že sa pýtate, ale v skutočnosti ten jav, ktorý študujem, nerobí absolútne nič.“ (Argument vplyvu na kariéru nie je platný, ale hovorím ho, aby som ponechal ústupovú líniu.)

Chalmers kritizuje dualizmus hmoty na základe toho, že je ťažko vidieť, aká nová fyzikálna teória, ktorej nová hmota interaguje s bežnou hmotou, by mohla vysvetliť vedomie. Ale dualizmus vlastností má presne ten istý problém. Bez ohľadu na to, o akom druhu duálnej vlastnosti hovoríte, ako presne to vysvetľuje vedomie?

Keď Chalmers predpokladal vlastnosť navyše, ktorá *je* vedomie, *urobil* skok cez nevysvetliteľné. Ako pomôže jeho teórii tvrdiť navyše, že táto vlastnosť navyše *nemá žiaden účinok*? Prečo by nemohla jednoducho mať?

Keby som chcel byť protivný, bol by to správny čas dotiahnuť draka – spomenúť podobenstvo Carla Sagana o drakovi v garáži. „Mám draka v garáži.“ Super! Chcem ho vidieť, poďme na to! „Nemôžeš ho vidieť, je to neviditeľný drak.“ Dobre, chcem si ho teda vypočuť. „Prepáč, je to nepočuteľný drak.“ Chcem odmerať jeho produkciu kysličníka uhličitého. „Nedýcha.“ Hodím do vzduchu za vrece múky, aby som videl jeho obrysy. „Tento drak prepúšťa múku.“

Jedným z motívov robiť svoju teóriu nefalzifikovateľnou je, že sa hlboko vnútri bojíte podrobiť ju skúške. Sir Roger Penrose (fyzik) a Stuart Hameroff (neurológ) sú dualisti hmoty; myslia si, že sa v kvantových veciach deje niečo tajomné, že Everett sa mýli a že „kolaps vlnovej funkcie“ je fyzikálne skutočný, a že práve tam žije vedomie a tak kauzálne vyplýva na vaše pery, keď nahlas poviete: „Myslím, teda som.“ Keďže v toto verili, predpovedali, že sa neuróny budú chrániť pred dekoherenciou dosť dlho na to, aby si udržali makroskopické kvantové stavy.

Toto sa zatiaľ testuje, a zatiaľ to pre Penrosea nevyzerá ružovo...

...ale Penroseov základný prístup je vedecky úctyhodný. Možno nie bayesovský, ale stále v zásade zdravý. Prišiel s nezvyčajou hypotézou. Povedal, ako ju testovať. Išiel a naozaj ju skúsil otestovať.

Ako som raz povedal Stuartovi Hameroffovi: „Myslím si, že hypotéza, ktorú testujete, je celkom beznádejná, a že vaše pokusy *jednoznačne* hodno finančne podporiť. Aj keď nenájdete presne to, čo hľadáte, hľadáte na mieste, na ktorom nehľadá nikto iný, a mohli by ste nájsť niečo zaujímavé.“

Sprostým odmietnutím epifenomenalizmu by teda mohlo byť, že zombi-isti sa boja povedať, že vec vedomia môže mať *účinky*, pretože potom by vedci mohli ísť *hľadať* tieto vlastnosti navyše, a nenájsť ich.

Nemyslím si však, že toto je naozaj pravda pre Chalmersa. Keby Chalmersovi chýbala poctivosť, mohol si veci *výrazne* zjednodušiť.

(Ale pre prípad, že by Chalmer toho čítal a bál by sa falzifikácie, chcem podotknúť, že ak je epifenomenalizmus nepravdivý, potom *existuje* nejaké iné vysetlenie toho, čo voláme vedomie, a jedného dňa sa nájde, čím sa Chalmersova teória zosype; ak teda Chalmersovi záleží na jeho mieste v histórii, nemá motív podporovať epifenomenalizmus, pokiaľ si naozaj nemyslí, že je pravdivý.)

Chalmers je jeden z najfrustrujúcejších filozofov, akých poznám. Občas sa čudujem, či sa nehrá na „Porazený ateizmus“. Chalmers robí naozaj *presnú* analýzu... a potom v poslednej chvíli zatočí vľavo. Vysvetlí všetko, čo je na scenári sveta zombií nesprávne, a potom, keď už zredukoval celý argument na cimpr-campr, pokojne ho prijme.

Chalmers robí to isté, keď pokojne podrobne vysvetlí, v čom je problém, keď povieme, že naša vlastná viera vo vedomie je oprávnená, keď naše zombícke dvojčatá povedia presne to isté, z presne rovnakých dôvodov, a mýlia sa.

Podľa Chalmersovej teórie, keď samotný Chalmers hovorí, že verí vo vedomie, nemôže na to mať *kauzálne* právo; jeho názor nie je spôsobený samotným faktom. V neprítomnosti vedomia by Chalmers písal celkom rovnaké články z rovnakých dôvodov.

Podľa epifenomenalizmu, Chalmersovo tvrdenie, že verí v existenciu vedomia, nemožno zdôvodniť ako výsledok procesu, ktorý systematicky produkuje pravdivé názory, pretože zombícke dvojča píše tie isté články pomocou toho istého systematického procesu a mýli sa.

Chalmers to pripúšťa. Chalmers v skutočnosti vysvetľuje tento argument do veľkej hĺbky vo svojej knihe. Dobre, takže Chalmers solídne dokázal, že nie je oprávnený veriť v epifenomenálne vedomie, hej? Nie. Chalmers píše:

> Vedomá skúsenosť spočíva v strede nášho epistemického vesmíru; máme k nej *priamy* prístup. Toto vyvoláva otázku: čo je to, čo oprávňuje našu vieru v naše skúsenosti, ak to nie je kauzálne spojenie s týmito skúsenosťami, a ak to nie je mechanizmus, ktorým sa táto viery vytvorila? Myslím si, že odpoveď na toto je jasná: to, že *máme* tieto skúsenosti, oprávňuje túto vieru. Napríklad, samotný fakt, že mám teraz zážitok červenej, poskytuje oprávnenie môjmu názoru, že mám zážitok červenej...
>
> Pretože moje zombícke dvojča nemá tento zážitok, je vo veľmi odlišnej epistemickej situácii odo mňa, a jeho úsudok nemá zodpovedajúce oprávnenie. Môže byť lákavé namietať, že ak sa môj názor nachádza v ríši fyziky, aj jeho oprávnenie sa musí nachádzať v ríši fyziky; ale to je *non sequitur*. Z faktu, že v ríši fyziky neexistuje oprávnenie, možno dôjsť k záveru, že *fyzikálna* časť mňa (povedzme môj mozog) nie je oprávnená tomuto veriť. Ale otázka je, či *ja* mám na tento názor oprávnenie, nie či ho má môj *mozog*, a ak je dualizmus vlastností pravdivý, potom ja som viac než iba môj mozog.

Takže – ak som túto tézu pochopil správne – existuje tvoje jadro, mimo tvojho mozgu, ktoré verí, že nie je zombia, a priamo zažíva, že nie je zombia; a tak sú jeho názory oprávnené.

Ale Chalmers práve *toto všetko napísal*, do svojej celkom fyzikálnej *knihy*, a to isté urobil aj zombík Chalmers.

Zombík Chalmers nemohol napísať túto knihu *preto*, lebo jeho zombícke jadro ja je nad jeho mozgom; musí na to byť nejaký celkom iný dôvod, v rámci zákonov fyziky.

Z toho vyplýva, že aj keď *existuje* skrytá časť Chalmersa, ktorá je vedomie a verí vo vedomie, priamo a nesprostredkovane, existuje aj *oddelený podpriestor* Chalmersa – kauzálne uzavretý kognitívny podsystém, ktorý koná celkom *v rámci* fyziky – a toto „vonkajšie ja“ je to, čo hovorí Chalmersovo vnútorné rozprávanie, a píše články o vedomí.

Nevidím žiaden spôsob, ako sa vyhnúť obvineniu, že podľa Chalmersovej vlastnej teórie, je tento oddelený vonkajší Chalmers pomätený. Toto je tá časť Chalmersa, ktorá je rovnaká v tomto svete i vo svete zombií; a ani v jednom z tých svetov nepíše filozofické články o vedomí *z rozumného dôvodu*. Chalmersove filozofické články nie sú výstupom jeho vnútorného jadra vedomia a viery vo vedomie, sú výstupom púhej fyziky vnútorného rozprávania, ktoré spôsobuje, že Chalmersove prsty stláčajú klávesy na jeho počítači.

A predsa tento pomätený vonkajší Chalmers píše filozofické články, ktoré sú *zhodou okolností* dokonale správne, *čo je nezávislý a dodatočný zázrak.* Nie je to logicky nevyhnutný zázrak (potom by svet zombií nebol logicky možný). Fyzikálne závislý zázrak, ktorý je zhodou okolnosti pravdivý v tom, čo považuje za náš vesmír, hoci veda nikdy nemôže odlíšiť náš vesmír od sveta zombií.

Prinajmenšom sa zdá, že toto by bol dôsledok toho, čo nám hovorí vonkajší Chalmers, ktorý je podľa vlastného tvrdenia pomätený.

Myslím si, že hovorím za všetkých redukcionistov, keď poviem: *He?*

Toto nie sú epicykly. Toto je: „Planéty sa pohybujú podľa týchto epicyklov – ale samotné epicykly v skutočnosti *nerobia* nič – existuje niečo iné, čo spôsobuje, že sa planéty pohybujú presne tak, ako by sa mali pohybovať podľa epicyklov, na čo nemám žiadne vysvetlenie – a mimochodom, toto isté by som tvrdil, aj keby žiadne epicykly neexistovali."

Mám neštandardný pohľad na filozofiu, pretože sa na všetko pozerám z uhla dizajnu UI; konkrétne sebamodifikujúcej sa Všeobecnej Umelej Inteligencie so stabilnou štruktúrou motivácie.

Keď myslím na dizajn UI, uvažujem nad princípmi ako je teória pravdepodobnosti, bayesovský pojem indície ako diferenciálnej diagnostiky, a najmä reflexívna koherencia. Ľubovoľná sebamodifikujúca UI, ktorá začne v reflexívne nekonzistentnom stave, tak dlho nevydrží.

Ak sa sebamodifikujúca UI pozrie na časť seba, ktorá usudzuje „B" v prípade A – časť seba, ktorá napíše do pamäte „B" vždy vtedy, keď platí podmienka A – UI túto časť preskúma, zistí, ako (kauzálne) funguje v kontexte širšieho vesmíru, a usúdi, že táto časť má sklon systematicky zapisovať do pamäte nepravdivé údaje, potom UI našla niečo, čo vyzerá ako chyba, a upraví sama seba, aby v prípade A nezapisovalo do zoznamu svojich názorov „B".

Ľubovoľná epistemická teória, ktorá neberie ohľad na reflexívnu koherenciu, nie je dobrá teória na použitie v sebazdokonaľujúcej sa UI. Toto je z môjho pohľadu pádny argument, keď vezmem do úvahy, *na* čo naozaj chcem filozofiu použiť. Musím teda aj tak vymyslieť reflexívne koherentnú teóriu. A keď to robím, čuduj sa svete, reflexívna koherencia začne dávať intuitívny zmysel.

Toto je teda zvyčajný spôsob, akým mám sklon myslieť na tieto veci. A teraz sa pozriem späť na Chalmersa:

Kauzálne uzavretý „vonkajší Chalmers" (ktorého nijakým spôsobom neovplyvňuje „vnútorný Chalmers", ktorý má svoje nezávislé dodatočné vedomie a názory) musí vykonávať nejakú systematicky nespoľahlivú, nezaručenú operáciu, ktorá *nejakým nevysvetleným spôsobom* spôsobuje, že vnútorný rozhovor vytvára názory o „vnútornom Chalmersovi", ktoré sú *správne bez akéhokoľvek logického dôvodu*, ktorý by existoval v našom vesmíre.

Nemôže existovať žiaden dôvod, aby vonkajší Chalmers *alebo ľubovoľná koherentná sebaskúmajúca UI* verili v túto tajomnú správnosť. Dobrý dizajn UI by mal podľa mňa vyzerať ako reflexívne koherentná inteligencia stelesnená v kauzálnom systéme, s *testovateľnou* teóriou o tom, ako tento samotný kauzálny systém vytvára systematicky presné názory na ceste k dosahovaniu svojich cieľov.

Preto UI preskenuje Chalmersa a uvidí uzavretý kauzálny poznávací systém vytvárajúci vnútorné rozprávanie, ktoré hovorí nezmysly. Nezmysly, ktoré zrejme majú veľký dopad na to, čo by sa Chalmersa *malo považovať za morálne hodnotnú osobu*.

Toto nemusí byť pre teoretikov UI *nevyhnutne* problém. Je to problém *iba* ak ste zhodou okolností epifenomenalista. Ak veríte buď redukcionistom (vedomie sa deje *v* atómoch) alebo dualistom hmoty (vedomie je *kauzálne schopná* nehmotná vec), ľudia, ktorí hovoria o vedomí, hovoria o niečom skutočnom, a reflexívne konzistentná bayesovská UI to môže vidieť stopovaním reťaze kauzality naspäť k tomu, čo spôsobuje, že ľudia hovoria „vedomie".

Podľa Chalmersa, kauzálne uzavretý poznávací systém Chalmersovho vnútorného rozprávania je (záhadne) pokazený spôsobom, ktorý, nie nevyhnutne, ale akurát v *našom* vesmíre, je zázračnou zhodou okolností správny. Navyše, toto vnútorné rozprávanie tvrdí, že „toto vnútorné rozprávanie je záhadne pokazené, ale zázračnou zhodou okolností správne odráža oprávnené myšlienky o svojom epifenomenálnom vnútornom jadre", a opäť, v *našom* vesmíre to je zázračnou zhodou okolností pravda.

Ale prosím vás!

Nemal by niekedy prísť moment, keď sa nejakej myšlienky jednoducho vzdáte? Kde si na nejakej holej intuitívnej úrovni skrátka poviete: *Čo som si to len myslel?*

Ľudstvo zhromaždilo nejaké rozsiahle skúsenosti s tým, ako vyzerajú správne teórie o svete. *Správna teória takto nevyzerá.*

„Argument nedôverčivosťou," poviete. Fajn, chcete to mať vyhláskované? Uvedená Chalmersova teória predpokladá viaceré nevysvetlené zložité zázraky. To znižuje jej apriórnu pravdepodobnosť podľa

pravidla konjunkcie pravdepodobností a Occamovej britvy. Preto jej dominujú prinajmenšom dve teórie, ktoré predpokladajú menej zázrakov, konkrétne:

- Dualizmus hmoty:
    - Existuje vec vedomia, ktorej zatiaľ nerozumieme, výnimočne nadfyzikálna vec, ktorá *viditeľne ovplyvňuje* náš svet; a táto vec spôsobuje, že hovoríme o vedomí.
- Redukcionizmus založený nie celkom na viere:
    - To, čo nazývame „vedomie“ sa deje *v rámci* fyziky, zatiaľ neznámym spôsobom, podobne ako keď za posledných tritisíc rokov ľudstvo narazilo na niečo tajomné.
    - Vaša intuícia, že žiadna hmota nemôže dávať dokopy vedomie, je nesprávna. Keby ste *naozaj* vedeli, *presne* prečo hovoríte o vedomí, dalo by vám to nové vhľady, ktorých tvar teraz neviete odhadnúť; a potom by ste si uvedomili, že vaše argumenty o tom, že v normálnej fyzike nie je miesto na vedomie, boli pomýlené.

Porovnajte s týmto:

- Epifenomenálny dualizmus vlastností:
    - Hmota má dodatočné vlastnosti vedomia, ktorým zatiaľ nerozumieme. Tieto vlastnosti sú epifenomenálne s ohľadom na bežnú pozorovateľnú fyziku – nespôsobujú žiaden rozdiel v pohybe častíc.
    - *Nezávisle* na tom existuje zatiaľ neznámy dôvod *v rámci normálnej fyziky*, prečo filozofi hovoria o vedomí a vymýšľajú teórie duálnych vlastností.
    - *Zázrakom*, keď filozofi hovoria o vedomí, premosťujúce zákony *nášho* sveta sú presne také, aby bolo toto hovorenie o vedomí správne, hoci vyplýva z nefungovania (vyvodzuje logicky nepodložené závery) v kauzálne uzavrenom kognitívnom systéme, ktorý píše filozofické články.

Viem, že tu hovorím z obmedzenej skúsenosti. Ale podľa mojej obmedzenej skúsenosti by argument zombií mohol byť kandidátom na *najpomätenejšiu myšlienku v celej filozofii*.

Sú chvíle, keď ako racionalista musíte veriť veciam, ktoré vám pripadajú čudné. Relativita vyzerá čudne, kvantová mechanika vyzerá čudne, prirodzený výber vyzerá čudne.

Ale všetka táto čudesnosť je podoprená masou dôkazou. Je rozdiel, či veríte niečomu čudnému, pretože veda to drvivo potvrdila...

...a keď veríte predpokladu, ktorý vyzerá celkom *pomätene*, pretože je to veľmi zložitý filozofický argument postavený okolo neurčitých zázrakov a veľkých prázdnych miest, o ktorých ani netvrdí, že im rozumie...

...v prípade, keď *ešte aj po prijatí všetkého, čo vám doteraz povedali*, bude daný jav aj potom vyzerať ako tajomstvo a stále bude mať tú istú vlastnosť úžasnej nepreniknuteľnosti, ktorú mal na začiatku.

Správna vec, ktorú by mal racionalista v tomto bode povedať, keby mu všetky argumenty Davida Chalmersa pripadali jednotlivo uveriteľné – čo mne nepripadajú – je:

„Dobre... nerozumiem, ako funguje vedomie... priznávam... a možno idem na celý problém z nesprávnej strany, alebo si kladiem nesprávne otázky... ale táto vec so zombiami *nemá šancu byť správna*. Tieto argumenty nie sú dosť podoprené, aby som im uveril – najmä keď by som po uverení nebol o nič menej zmätený. Na inštinktívnej úrovni, toto jednoducho *nevyzerá* ako spôsob, ktorým by skutočnosť mohla *naozaj naozaj* fungovať.“

Pozor, že netvrdím, že toto je náhrada za podrobné analytické vyvrátenie Chalmersovej tézy. Systém 1 nie je náhradou za systém 2, hoci môže pomôcť ukázať cestu. Stále potrebujete vystopovať, v čom *konkrétne* je problém.

Chalmers napísal veľkú knihu, ktorá nie je celá dostupná cez voľný náhľad googlu. Nezopakoval som dlhé reťaze argumentov, kde Chalmers pokojne podrobne vykladá argumenty proti sebe. Pokúsil som sa iba o záverečné vyvrátenie Chalmersovej naposledy predloženej obhajoby, na ktoré Chalmers podľa mojich vedomostí zatiaľ nereagoval. Akoby som tým odrazil loptu na jeho stranu ihriska.

Ale áno, v jadre *príčetná* vec, ktorú treba urobiť, keď vidíte záver argumentu zombií, je povedať: „Toto *nemôže* byť pravda“ a začať hľadať chybu.

*→

222. Reakcia na zombie

Dnes som trochu unavený, lebo som do tretej ráno písal včerajší vyše 6000-slovný článok o zombiách, takže odpoviem iba Richardovi, a zaplním jednu medzeru, ktorú som si všimol na druhý deň.

(A) Richard Chappell píše:

> Terminologická poznámka (aby sme sa vyhli zbytočnému nedorozumeniu): to, čo nazývaš „predstaviteľné“, by sme my ostatní nazvali iba „*napohľad* predstaviteľné“.

Medzera medzi „nevidím rozpor“ a „toto je logicky možné“ je taká veľká (NP-úplná dokonca aj v niektorých jednoducho vyzerajúcich prípadoch), že by sme naozaj mali mať dve rôzne slová. Keďže argument zombií je taký zložitý, že túto obrovskú medzeru možno zamiesť pod koberec drobných terminologických rozdielov, naozaj si myslím, že by bol dobrý nápad hovoriť „predstaviteľné“ verzus „logicky možné“ alebo možno mať ešte viditeľnejšie rozlíšenie. Nemôžem zmeniť už zavedenú profesionálnu terminológiu, ale v prípadoch, ako je tento, môžem vážne odmietnuť používať ju.

Mohol by som používať „napohľad predstaviteľné“ pre ten druh informácií, ktoré zástancovia zombií dostávajú, keď si predstavujú svety zombií, a „logicky možné“ pre ten druh informácií, ktoré sú potvrdené ukázaním úplného modelu alebo logického dôkazu. Všimnite si veľkosť medzery medzi informáciou, ktorú dostanete, keď zavriete oči a predstavíte si zombie, a informáciou, ktorou potrebujete podložiť argument za epifenomenalizmus.

> Čiže tvoj pohľad by sa dal charakterizovať ako materializmus typu A, názor, že zombie nie sú ani (naozaj) predstaviteľné, nieto ešte metafyzicky možné.

Materializmus typu A je veľký balík; nemali by ste mi pripisovať celý balík, kým ma nevidíte súhlasiť s každou jeho časťou. Myslím si, že keď sa niekto opýta „Čo je to vedomie?“, kladie legitímnu otázku, a má legitímne právo na vhľad; nemyslím si, že *odpoveď* musí mať nutne tvar „Tu je táto vec, ktorá má všetky vlastnosti, ktoré by si pripisoval vedomiu, z takých a onakých dôvodov“, ale mala by sa do istej miery skladať z vhľadov, ktoré spôsobia, že si uvedomíte, že ste túto otázku zle kládli.

Nejde tu o eliminácii vedomia. Ide tu o realistickosť ohľadom druhu vhľadov, ktoré môžeme očakávať, keď čelíme problému, ktorý (1) vyzerá, že musí mať *nejaké* riešenie, (2) vyzerá, že nemôže mať žiadne riešenie, a (3) *diskutujeme* o ňom spôsobom, ktorý veľmi závisí od nie celkom pochopenej ad-hoc architektúry ľudského poznania.

> (1) Pokiaľ môžem povedať, zatiaľ si neidentifikoval žiaden *logický rozpor* v opise sveta zombií. Iba si ukázal, že je to akosi zvláštne. Ale existujú mnohé bizarné možné svety. To nie je dôvod tvrdiť implicitný spor. Takže stále je mi celkom záhadou, čo má byť tento domnelý rozpor.

Okej, vysvetlím to z materialistického pohľadu:

1. Svet zombií podľa definície obsahuje všetky časti nášho sveta, ktoré sú v uzávere vzťahov „spôsobený týmto“ alebo „účinok tohto“ ľubovoľného pozorovateľného javu. Konkrétne obsahuje *príčinu* môjho viditeľného povedania: „Myslím, teda som.“
2. Keď sústredím svoju vnútornú pozornosť na svoju vnútornú pozornosť, krátko nato vnímam, že moje vnútorné rozprávanie hovorí: „Sústreďujem svoju vnútornú pozornosť na svoju vnútornú pozornosť“, a môžem to povedať nahlas, ak chcem.
3. Intuitívne mi pripadá isté, že moja vnútorná pozornosť spôsobila, že moje vnútorné rozprávanie povedalo určité veci, a že moje vnútorné rozprávanie spôsobilo, že moje pery povedal určité veci.

→ http://lesswrong.com/lw/p7/zombies_zombies/

4. Slovo „vedomie“, ak vôbec niečo znamená, odkazuje na to-čo-je, alebo to-čo-spôsobuje, alebo to-čo-spôsobuje-že-hovorím-že-mám vnútornú pozornosť.

5. Z (3) a (4) by teda vyplývalo, že ak je svet zombií uzavretý vzhľadom na príčiny môjho povedania „myslím si, teda som“, potom svet zombií obsahuje to, na čo ja odkazujem ako „vedomie“.

6. Podľa definície, svet zombií neobsahuje vedomie.

7. (3) je podľa mňa s veľmi vysokou pravdepodobnosťou empiricky pravdivé. Preto vyhodnocujem ako vysoko empiricky pravdepodobné, že svet zombií je logicky nemožný.

Svet zombií môžete zachrániť tým, že necháte príčinu toho, že moje vnútorného rozprávanie hovorí „myslím, teda som“, byť niečím celkom iným ako vedomím. V spojení s predpokladom, že vedomie existuje, mi táto časť pripadá pomätená.

Ale ak si vieme *predstaviť* horeuvedené, nie je svet zombií predstaviteľný?

Nie, pretože dve rôzne konštrukcie sveta zombií by zahŕňali dať slovu „vedomie“ rôzne empirické referenty, ako keď slovo „voda“ v našom svete znamená H2O a slovo „voda“ v Putnamovej dvojčati Zeme znamená XYZ. Aby bol svet zombií logicky možný, nestačí, aby vzhľadom na tom, čo *vy* viete o fungovaní empirického sveta, slovo „vedomie“ *mohlo* odkazovať na epifenomén, ktorý je celkom iný ako to vedomie, ktoré poznáme. Svetu zombií chýba vedomie, nie „vedomie“ - je to svet bez H20, nie svet bez „vody“. Toto je potrebné na podporenie empirického tvrdenia: „Môžete odstrániť referent alebo čokoľvek sa myslí slovom ‚vedomie‘ z nášho sveta, zatiaľ čo ponecháte všetky atómy na tom istom mieste.“

Čo znamená: Považujem za *empirický* fakt, vzhľadom na to, na čo slovo „vedomie“ naozaj odkazuje, že je *logicky* nemožné odstrániť vedomie bez pohnutia nejakým atómom. Čo by znamenalo odstrániť zo sveta „vedomie“ a nie vedomie, o tom nebudem špekulovať.

> (2) Je zavádzajúce povedať, že je „zázračné“ (z pohľadu dualizmu vlastností), že naše kváliá sú v takej zhode s fyzikálnym svetom. Existuje predsa prírodný zákon, ktorý to zaručuje. Nie je to teda o nič viac zázračné, než ľubovoľná iná logicky podminená nevyhnutnosť v zákonoch (napríklad konštanty v našich fyzikálnych zákonoch).

Samotný tento prírodný zákon je „zázračný“ - počíta sa ako dodatočný zložitý a nepravdepodobný prvok predkladanej teórie, bez toho, že by bol sám zdôvodnený pomocou už známych vecí. Predpokladá sa (a) vnútorný svet, ktorý je vedomý, a (b) nefungujúci vonkajší svet, ktorý bezvôdovne hovorí o vedomí, (c) ktoré sú dokonale zladené. Výrok (c) nevyplýva z (a) a (b), je to teda nezávislý predpoklad.

Súhlasím, že toto použitie „zázraku“ odporuje filozofickému významu porušenia prírodného zákona; myslel som to v zmysle nepravdepodobnosti, ktorá neprichádza zo žiadneho zrejmého zdroja, ako názor typu perpetuum mobile. Preto bolo toto slovo v danom kontexte zle vybrané. Ale nie to *intuitívne* ten typ veci, ktorý by sme mali volať zázrakom? Vaše vedomie naozaj nespôsobuje, že hovoríte, že máte vedomie; existuje nezávislá fyzikálna vec, ktorá spôsobuje, že hovoríte, že máte vedomie; ale existuje aj zákon, ktorý tieto dve veci zarovnáva – toto je veru udalosť podobného rádu nezvyčajnosti ako že oplátka nadobúda podstatu Kristovho tela, zatiaľ čo má presný výzor a vonkajšie správanie oplátky, veď viete, skrátka existuje prírodný zákon, ktorý to zaručuje.

> To znamená, že zombícky (alebo „vonkajší“) Chalmers v skutočnosti k *ničomu* nedošiel, lebo jeho výroky sú nezmyselné. A fortiori, nedošiel k ničomu bezdôvodne. Skrátka vydáva zvuky; ktoré nepodliehajú epistemickému hodnoteniu o nič viac než vtáčie štebotanie.

Keď sa na toto pozriem z hľadiska návrhu UI, pripadá mi, že by ste mali vedieť navrhnúť UI, ktorá systematicky zdokonaľuje svoju vnútornú časť, ktorá koreluje (v zmysle vzájomnej informácie alebo systematickej korelácie) s prostredím, možno pomocou desatinných čísel toho typu, ktorý nazývam „pravdepodobnosti“, pretože ich vzájomné vzťahy sa riadia Coxovými vetami, keď UI získa nové informácie – pardon, nové zmyslové vstupy.

Môžete povedať, že pokiaľ UI nie je viac než púhe tranzistory – pokiaľ nemá ten duálny aspekt – že UI nemá názory.

Myslím si, že som svoj pohľad na toto vysvetlil pomerne jasne v kapitole „Jednoduchá pravda“.

Mne pripadá veľmi priamočiare vytvárať mapy, ktoré systematickým spôsobom korelujú s územím, bez spomínania hocičoho iného než vecí čisto fyzikálnej kauzality. UI vypíše mapu Texasu. Iná UI poletí s touto mapou do Texasu a skontroluje, či sú diaľnice na príslušných miestach, zapípa „Pravda“, keď nájde zhodu, a „Nepravda“, keď nájde nezhodu. Môžete odmietnuť natývať toto „mapa Texasu“, ale samotné UI budú stále pípať „Pravda“ alebo „Nepravda“, a spomínané UI zapípajú „Nepravda“, keď sa pozrú na Chalmersovu vieru v epifenomenálne vnútorné jadro, a ja by som s nimi veru súhlasil.

Je jasné, že *funkciu mapovania skutočnosti* vykonáva výhradne Vonkajší Chalmers. Celá záležitosť *vytvárania názorových reprezentácií* sa riadi bayesovskou štruktúrou v kauzálnych interakciách. Nezostáva nič, čo by mohol robiť Vnútorný Chalmers, okrem požehnania celej tejto veci epifenomenálnym *zmyslom*. Kde tento „zmysel“ je niečo celkom nesúvisiace so systematickou korešpondenciou mapy a územia, alebo so schopnosťou používať túto mapu na usmerňovanie skutočnosti. Keď teda začneme hovoriť o „presnosti“ a tobôž „systematickej presnosti“, zdá sa mi, že by sme ju mali vedieť určiť výhradne pozeraním na Vonkajšieho Chalmersa.

(B) Vo včerajšom texte som vynechal jeden predpoklad, keď som písal:

> Ak sa sebamodifikujúca UI pozrie na časť seba, ktorá dochádza k záveru „B“ v situácii A – časť seba, ktorá zapíše do pamäte „B“, kedykoľvek platí stav A – a UI túto časť preskúma, zistí ako (kauzálne) funguje v kontexte širšieho vesmíru, a dôjde k záveru, že táto časť systematicky zapisuje do pamäte nepravdivé údaje, potom UI našla niečo, čo vyzerá ako chyba, a UI sa zmení, aby nezapisovala „B“ do zbierky názorov za okolností A.
>
> …
>
> Vonkajší Chalmers *alebo ľubovoľná reflexívne koherentná sebaskúmajúca UI* však nemá žiaden možný dôvod veriť v túto tajomnú správnosť. Dobrý dizajn UI by podľa mňa bola reflexívne koherentná inteligencia s testovateľnou teóriou o tom, ako funguje ako kauzálny systém, a teda s testovateľnou teóriou o tom, ako tento kauzálny systém vytvára systematicky presné názory na ceste k dosahovaniu svojich cieľov

V skutočnosti potrebujeme k horeuvedenému ešte ďalší predpoklad, a ten je, že „dobrý dizajn UI“ (prinajmenšom ten druh, na ktorý myslím) posudzuje svoju rozumnosť modulárnym spôsobom; zabezpečuje globálnu rozumnosť zabezpečovaním lokálnej rozumnosti. Ak existuje časť, ktorá je relatívne voči svojmu kontextu lokálne systematicky nespoľahlivá – pre nejaké možné názory „B_i“ a podmienky A_i pridáva do lokálnej zbierky názorov „B_i“ za podmienky A_i, kde po reflexii systém ukazuje, že B_i nie je pravdivé (alebo v prípade pravdepodobnostných názorov, nie presné), ak je pravdivá lokálna podmienka A_i, potom toto je chyba. Tento druh modularity je *jeden* možný spôsob, ako urobiť problém zvládnuteľným, je to je spôsob, ktorý v súčasnosti považujem za dizajn prvej generácie UI. [Úprava 2013: Skutočná predstava, na ktorú som tu myslel, bola teraz rozpísaná a formalizovaná v článku Tiling Agents for Self-Modifying AI, sekcia 6.]

Tá predstava je, že kauzálne uzavretý poznávací systém – ako je UI navrhnutá svojimi programátormi, aby používala iba kauzálne pôsobiace časti; alebo UI, ktorej teória o vlastnom fungovaní je celá testovateľná; alebo vonkajší Chalmers, ktorý píše filozofické články – ktorý verí, že má epifenomenálne vnútorné ja, musí robiť niečo systematicky nespoľahlivé, pretože by došiel k rovnakému záveru aj vo svete zombií. Myseľ, ktorej všetky časti sú systematicky lokálne spoľahlivé, relatívne k svojmu kontextu, by bola systematicky globálne spoľahlivá. Teda myseľ, ktorá je globálne nespoľahlivá, musí obsahovať aspoň jednu lokálne nespoľahlivú časť. Takže kauzálne uzavretý kognitívny systém skúmajúci svoju vlastnú lokálnu spoľahlivosť musí objaviť, že aspoň jeden krok, ktorý sa zúčastnil na pridaní názoru o epifenomenálnom vnútornom ja, je nespoľahlivý.

Ak existujú iné spôsoby, ako môže byť myseľ reflexívne koherentná, ktoré sa vyhýbajú tomuto dôkazu neviery v zombie, nech sa filozofom páči skúsiť ich špecifikovať.

Toto všetkom musím špecifikovať preto, lebo inak by ste dostali nejaký druh extrémne lacnej reflexívnej koherencie, kde by UI nikdy neoznačila sama seba za nespoľahlivú. Napríklad keby UI našla v sebe súčiastku, ktorá počíta 2 + 2 = 5 (v kontexte počítania oviec), UI by uvažovala: „Nuž, táto súčiastka nefunguje a hovorí, že 2 + 2 = 5... ale čistou zhodou okolností 2 + 2 *je* 5, aspoň mi to tak pripadá... takže aj keď táto časť vyzerá systematicky nespoľahlivo, radšej ju nechám tak, ako je, inak bude tento špeciálny prípad vyhodnocovať zle.“ Pre toto hovorím o zabezpečovaní globálnej spoľahlivosti zabezpečovaním lokálnej systematickej spoľahlivosti – ak svoje globálne názory porovnávate iba so svojimi globálnymi názory, nikam sa nedostanete.

Z tohto je všeobecné ponaučenie: Ukážte, že vaše argumenty sú globálne spoľahlivé tak, že každý krok je lokálne spoľahlivý, neporovnávajte iba závery argumentov s vašimi intuíciami. [Úprava 2013: Pozrite si Proof, Implications, and Models [Dôkazy, implikácie a modely], kde sa preberá fakt, že platná logika je lokálne platná.]

(C) Anonymný pisateľ napísal:

> Toto je na okraj, ale myslím si, že tvoja etymológia pre „n'shama“ je nesprávna. Súvisí to so slovom „dýchať“, nie „počuť“. Slovný koreň pre počutie obsahuje ayin, ktorý n'shama nemá.

Tak tomuto hovorím zázračne zavádzajúca zhoda okolností – hoci slovo N'Shama vzniklo z celkom iných dôvodov, znie *presne tak*, že som si myslel, že odkazuje na vnútorného poslucháča.

Joj.

*→

223. Všeobecný antizombícky princíp

> Každý problém, ktorý vyriešim, sa stane pravidlom, ktoré neskôr slúži na riešenie iných problémov.
>
> --Rene Descartes, Rozprava o metóde[204]

„Zombie“ sú domnelé bytosti, ktoré sú atóm za atómom totožné s nami, riadené rovnakými fyzikálnymi zákonmi pozorovateľnými treťou stranou, akurát že nemajú vedomie.

Hoci je ich filozofia zložitá, argument proti zombiám je v jadre jednoduchý: Keď zameriate svoje vnútorné uvedomovanie na svoje vnútorné uvedomovanie, krátko potom vaše vnútorné rozprávanie (malý hlas vo vašej hlave, ktorý hovorí vaše myšlienky) povie: „uvedomujem si, že si uvedomujem“, a potom to poviete nahlas, a potom to napíšete na klávesnici počítača a vyvoríte článok na blogu viditeľný pre tretiu stranu.

Vedomie, nech už je to čokoľvek – hmota, proces, meno pre zmätok – nie je epifenomenálne; vaša myseľ dokáže prichytiť vnútorného poslucháča pri akte počúvania, a povedať to nahlas. *Skutočnosť, že som napísal tento odstavec*, prinajmenšom *vyzerá*, že by mala vyvrátiť myšlienku, že vedomie nemá žiadne experimentálne zistiteľné dôsledky.

Nerád pri takejto filozoficky kontroverznej otázke hovorím „a teraz to už odsúhlasme a poďme ďalej“, ale zdá sa mi, že prevažná väčšina diskutérov na *Overcoming Bias* s týmto súhlasí. A existujú ďalšie závery, ku ktorým sa môžete dostať až keď prijmete, že nemôžete odobrať vedomie a nechať celý vesmír vyzerať celkom rovnako. Prijmime to teda a poďme ďalej.

Tvar tohto argumentu proti zombiám vyzerá, že by sa mal dať zovšeobecniť, stať sa antizombíckym princípom. Ale čo je vhodné zovšeobecnenie?

Povedzme napríklad, že niekto povie: „Mám v ruke prepínač, ktorý žiadnym spôsobom neovplyvňuje tvoj mozog; a práve vtedy, keď tento prepínač prepnem, prestaneš mať vedomie.“ Vylučuje antizombícky princíp aj toto, pomocou rovnakej štruktúry argumentu?

→ http://lesswrong.com/lw/p8/zombie_responses/

204 René Descartes, *Discours de la Méthode*, vol. 45 (Librairie des Bibliophiles, 1887).

Zdá sa mi, že v uvedenom prípade je odpoveď áno. Konkrétne môžete povedať: „Aj po prepnutí prepínača budem hovoriť o vedomí *z presne rovnakých dôvodov* ako predtým. Ak mám vedomie teraz, budem mať vedomie aj po prepnutí prepínača."

Filozofi môžu namietať: „Ale teraz kladieš rovnosť medzi vedomie a hovorenie o vedomí! Čo vládca zombií, debatný robot, ktorý opakuje premiešaný korpus amatérskej ľudskej diskusie o vedomí?"

Lenže ja som *nedal* rovnosť medzi „vedomie" a slovné správanie. Predpoklad v jadre je, že *okrem iných vecí*, skutočný referent „vedomia" je *tiež príčinou* prečo *ľudia* hovoria o vnútorných poslucháčoch.

Ako som tvrdil (dosť podrobne) v postupnosti o slovách, nie vždy chcete pri definovaní slova dokonalú aristotelovskú definíciu nutného a postačujúceho; niekedy chcete iba mapu pokladu, ktorá vás vedie k extenzionálnemu referentu. Takže „to, čo *naozaj* spôsobuje, že hovorím o nevysloviteľnom uvedomovaní" nie je nutná a postačujúca definícia. Ale ak to, čo *naozaj* spôsobuje, že hovorím o nevysloviteľnom uvedomovaní, nie je „vedomie", potom...

…potom je táto debata fakt márna. To nie je zdrvujúci argument proti zombiám – empirickú otázku nemožno vyriešiť púhou zložitosťou debaty. Ale ak sa pokúsite vzdorovať antizombíckemu princípu, budete mať problémy so *zmyslom* svojich tvrdení, nie iba s ich uveriteľnosťou.

Mohli by sme *definovať* slovo „vedomie" ako „to, čo naozaj spôsobuje, že ľudia hovoria o ‚vedomí'"? Toto by malo tú silnú výhodu, že by sme vedeli, že existuje aspoň jeden skutočný fakt označovaný slovom „vedomie". Dokonca aj keby naša viera vo vedomie bola zmätok, „vedomie" by pomenovávalo tú poznávaciu architektúru, ktorá tento zmätok vytvorila. Ale stanoviť definíciu je iba sľúbiť, že budeme nejaké slovo používať konzistentne; nerieši to žiadne empirické otázky, ako napríklad či naše vnútorné uvedomovanie spôsobuje, že hovoríme o svojom vnútornom uvedomovaní.

Vráťme sa k tomu prepínaču.

Keby sme dovolili, aby sa antizombícky argument vzťahoval aj na ten prepínač, potom by všeobecný antizombícky princíp *nehovoril* iba: „Ľubovoľná zmena, ktorá nie je principiálne experimentálne detekovateľná (PED), nemôže odstrániť vaše vedomie." Prepnutie prepínača je experimentálne detekovateľné, ale stále vyzerá *veľmi* nepravdepodobne, že by odstránilo vaše vedomie.

Možno antizombícky princíp hovorí: „Ľubovoľná zmena, ktorá vás žiadnym PED spôsobom neovplyvňuje, nemôže odstrániť vaše vedomie"?

Je však rozumné tvrdiť, že prepnutie prepínača vás neovplyvňuje *žiadnym* PED spôsobom? Všetky častice v prepínači interagujú s časticami, z ktorých sa skladá vaše telo a mozog. Existujú gravitačné účinky – maličké, ale skutočné a PED. Gravitačný ťah z jedného gramu prepínača vzdialeného desať metrov je približne 6×10^{-16} m/s^2. To je zhruba polovica priemeru neutrónu za sekundu za sekundu, čo je menej ako tepelný šum, ale omnoho viac než Planckova úroveň.

Mohli by sme prepnúť niekoľko svetelných rokov vzdialený prepínač, a v tom prípade by prepnutie na vás nemalo žiaden okamžitý kauzálny účinok (nech už v tomto prípade „okamžitý" znamená čokoľvek) (ak je štandardný model fyziky správny).

Ale nezdá sa, že by sme *museli* zmeniť tento myšlienkový experiment týmto spôsobom. Zdá sa, že ak sa prepne oddelený prepínač na opačnej strane miestnosti, nemali by ste očakávať, že váš vnútorný poslucháč zhasne ako svieca, pretože ten prepínač „samozrejme nemení" to, čo je skutočnou príčinou, že hovoríte o vnútornom poslucháčovi. Čokoľvek naozaj ste, nepredpokladáte, že to prepínač ovplyvní.

Toto je *veľký* krok.

Ak popierate, že je to rozumný krok, potom by ste radšej nikdy nemali chodiť blízko prepínačov. Ale aj tak, je to veľký krok.

Kľúčová myšlienka redukcionizmu je, že naše mapy vesmíru sú viacúrovňové, aby sme šetrili výpočtovú silu, ale fyzika vyzerá byť výhradne jednoúrovňová. Všetky naše diskusie o vesmíre sa odohrávajú pomocou referentov vysoko nad úrovňou základných častíc.

Prepnutie prepínača *zmení* základné častice vo vašom tele a mozgu. Posunie ich to o celý premer neutrónu odtiaľ, kde by inak boli.

V bežnom živote takúto malú zmenu odbijeme slovami, že prepínač „vás neovplyvňuje“. Lenže on vás *ovplyvňuje*. Mení to všetko o celé priemery neutrónov! Čo by mohlo zostať rovnaké? Jedine *popisy*, ktoré by ste dali na vyšších úrovniach organizácie – bunky, bielkoviny, signály putujúce neurónovým axónom. Pretože mapa je omnoho menej podrobná ako územie, musí mapovať mnoho rôznych stavov na rovnaký popis.

Ľubovoľný rozumný ľudsky vyzerajúci *popis* mozgu, ktorý hovorí o neurónoch a vzoroch aktivity (alebo dokonca o stavbe jednotlivých mikrotubulov, z ktorých sa axóny a dendrity skladajú) sa nezmení, keď prepnete prepínač na opačnej strane miestnosti. Jadrá sú väčšie než neutróny, atómy sú väčšie než jadrá, a kým sa dostaneme k rozprávaniu o *molekulárnej* úrovni, tá maličká gravitačná sila zmizla zo zoznamu vecí, ktoré sa unúvate *sledovať*.

Ale ak budete pridávať dosť maličkých gravitačných ťahov, jedného dňa vás hodia krížom cez miestnosť a roztrhajú na kúsky slapovými silami, takže očividne malý účinok *neznamená* „žiaden účinok“.

Možno tá slapová sila z toho maličkého potiahnutia *úžasnou* zhodou okolností potiahne jeden ión vápnika len o kúsok bližšie k iónovému kanálu a spôsobí, že bude vtiahnutý o kúsoček skôr, čo spôsobí, že jeden neurón vyšle signál mikroskopicky skôr než by to bol urobil inak, tento rozdiel sa chaoticky zosilní, a nakoniec spôsobí celý neurónovú vlnu, ktorá by inak nebola nastala, čím vás navedie na inú myšlienku, ktorá spôsobí epileptický záchvat, ktorý vás zabije, čo spôsobí, že prestanete mať vedomie...

Ak pospájate veľa maličkých kvantitatívnych účinkov, dostanete veľký kvalitatívny účinok – dosť veľký na to, aby pohol hocičím, čo sa unúvate pomenovať. Preto tvrdiť, že prepínač má doslova *nulový* účinok na veci, ktoré vás zaujímajú, znamená preháňať.

Ale pri jedinom vypínači je vyvinutá sila omnoho menšia než tepelná neistota, tobôž kvantová neistota. Ak neočakávate, že vaše vedomie bude blikavo vznikať a zanikať v dôsledku tepelného kmitania, potom by ste určite nemali očakávať, že zhasne ako svieca, keď si niekto o kilometer ďalej kýchne.

Bdelý bayesovec si všimne, že som práve urobil argument o *očakávaniach*, stavoch *poznania*, oprávnených *názoroch* na to, čo môže a čo nemôže vypnúť vaše vedomie.

Toto nemusí nevyhnutne zničiť antizombícky argument. Pravdepodobnosti nie sú istoty, ale *zákony* pravdepodobnosti sú zákony; ak rozumnosť hovorí, že niečomu nemôžete veriť na základe vašej terajšej informácie, potom je to zákon, nie odporúčanie.

Aj tak je táto verzia antizombíckeho argumentu slabšia. Nemá to pekné, čisté, absolútne jasne dané postavenie tvrdenia: „Nemôžete odstrániť vedomie, ak ponecháte všetky atómy na *presne* rovnakom mieste.“ (Alebo namiesto „všetky atómy“ dosaďte „všetky príčiny s principiálne experimentálne detekovateľnými účinkami“, a „rovnaká vlnová funkcia“ namiesto „rovnaké miesto“, atď.)

Ale nová verzia antizombíckeho argumentu stále platí. Môžete povedať: „Neviem, čo je vedomie naozaj je, a mám podozrenie, že môžem byť v tejto otázke zásadne popletený. Ale ak toto slovo vôbec niečo označuje, označuje to niečo, čo je okrem iných vecí príčinou môjho hovorenia o vedomí. Ja teda neviem, prečo hovorím o vedomí. Ale odohráva sa to vnútri mojej lebky a očakávam, že to nejako súvisí so signálmi neurónov. Alebo možno keby som vedomiu naozaj rozumel, musel by som hooriť o ešte základnejšej úrovni, napríklad o mikrotubuloch alebo neurotransmiteroch šíriacich sa synaptickými kanálmi. Každopádne, ten prepínač, ktorý si práve prepol, má na moje neurotransmitery a mikrotubuly omnoho menší vplyv než je tepelný šum pri 310 kelvinoch. Takže nech je skutočnou príčinou môjho rozprávania o vedomí čokoľvek, neočakávam, že to bude výrazne ovplyvnené gravitačným ťahom z toho prepínača. Mohlo by to byť mikroskopicko máličko ovplyvnené? Ale určite to nezhasne ako svieca. Očakávam, že potom budem ďalej hovoriť o vedomí *takmer presne* rovnakým spôsobom z *takmer presne* rovnakých dôvodov.“

Toto použitie antizombíckeho princípu je slabšie. Ale je omnoho všeobecnejšie. A v zmysle číreho zdravého rozumu, je správne.

Redukcionista a dualista hmoty v skutočnosti majú dve rôzne verzie horeuvedeného tvrdenia. Redukcionista navyše povie: „Čokoľvek spôsobuje, že hovorím o vedomí, zdá sa, že tie dôležité časti sa

odohrávajú na omnoho vyššej funkčnej úrovni než sú jadrá atómov. Niekto, kto rozumie vedomiu, by dokázal abstrahovať od jednotlivých signálov neurónov a hovoriť o poznávacej architektúre na vyššej úrovni, a stále by opisoval, ako moja myseľ tvorí myšlienky ako ‚myslím, teda som'. Posúvať veci o priemer atómového jadra by teda nemalo ovplyvniť moje vedomie (nanajvýš možno s veľmi malou pravdepodobnosťou, alebo o veľmi maličké množstvo, alebo až po významnom oneskorení)."

Dualista hmoty navyše povie: „Čokoľvek spôsobuje, že hovorím o vedomí, musí to byť niečo mimo výpočtovej fyziky, ktorú poznáme, čo znamená, že by to mohlo zahŕňať kvantové javy. Ale aj tak, moje vedomie nebliká vždy keď si niekto o kilometer ďalej kýchne. Keby blikalo, *všimol* by som si to. Bolo by to ako preskočiť pár sekúnd, alebo sa prebrať z umelého spánku, alebo by niekto povedal: ‚ja nemyslím, teda nie som'. Keďže je fyzikálny fakt, že tepelné vibrácie neovplyvňujú hmotu môjho vedomia, neočakávam ani, že ju naruší prepnutie prepínača."

Každopádne by ste *nemali* očakávať, že váš zmysel pre uvedomovanie zanikne, keď niekto povie slovo „Abrakadabra", aj keď to má nejaký mikroskopický fyzikálny účinok na váš mozog...

Ale počkate! Ak *počujete*, že niekto povie slovo „Abrakadabra", má to veľmi viditeľný účinok na váš mozog – taký veľký, že si to váš mozog dokáže všimnúť. Môže to zmeniť vaše vnútorné rozprávanie; môžete si pomyslieť: „Prečo ten človek práve povedal: ‚Abrakadabra'?"

Áno, ale *stále* očakávate, že potom budete ďalej hovoriť o vedomí takmer presne rovnakým spôsobom, z takmer presne rovnakých dôvodov.

A opäť, nie je to o tom, že by sme „vedomie" *stotožnili* s „tým, čo spôsobuje, že hovoríte o vedomí". Iba toľko, že vedomie, *okrem iných vecí*, spôsobuje, že hovoríte o vedomí. Takže hocičo, čo spôsobí, že vaše vedomie zhasne ako svieca, by malo spôsobiť, že prestanete hovoriť o vedomí.

Ak vám urobíme niečo, kde nevidíte, ako by to *mohlo* zmeniť vaše vnútorné rozprávanie – ten malý hlas vo vašej hlave, ktorý občas hovorí veci ako „myslím, teda som", a tie slová sa môžete rozhodnúť povedať nahlas – potom by to nemalo spôsobiť, že prestanete mať vedomie.

A toto je pravda dokonca aj keď je to vnútorné rozprávanie „prakticky rovnaké", a keď sú jeho príčiny prakticky rovnaké; medzi tými prakticky rovnakými príčinami je aj to, čo nazývate „vedomie".

Ak sa čudujete, kam toto celé smeruje a prečo je také dôležité tak dlho rozoberať taký napohľad zrejmý všeobecný antizombícky princíp, potom si predstavte nasledujúcu debatu:

> Albert: „Predstav si, že by som nahradil všetky neuróny v tvojom mozgu maličkými robotickými umelými neurónmi, ktoré by mali rovnaké spojenia, rovnaké lokálne vstupno-výstupné správanie, a analogický vnútorný stav a učiace sa pravidlá."
>
> Bernice: „To by ma zabilo! Už by neexistovala bytosť s vedomím."
>
> Charles: „No, existovala by bytosť s vedomím, ale nebol by som to *ja*."
>
> Sir Roger Penrose: „Myšlienkový pokus, ktorý navrhuješ, je nemožný. *Nemôžeš* duplikovať správanie neurónov, lebo narazíš na kvantovú gravitáciu. Keď som to povedal, nemá zmysel, aby som sa ďalej zúčastňoval tejto konverzácie." *(Odchádza preč.)*
>
> Albert: „Predstavme si, že by sa nahrádzanie robilo jeden neurón po druhom, a že by výmena nastala tak rýchlo, že by to nespôsobilo žiaden rozdiel v globálnom spracovaní."
>
> Bernice: „A to by bolo ako možné?"
>
> Albert: „Malý robot pripláva k neurónu, oblapí ho, preskenuje, naučí sa ho duplikovať, a potom náhle prevezme jeho správanie, medzi dvoma vlnami. V skutočnosti je táto napodobenina *taká* dobrá, že tvoje vonkajšie správanie je celkom rovnaké ako by bolo, keby sme mozog ponechali nedotknutý. Možno nie *presne* rovnaké, ale kauzálny dopad je omnoho menší než tepelný šum pri 310 kelvinoch."
>
> Charles: „No a čo?"

Albert: „Neporušujú tvoje názory všeobecný antizombícky princíp? Čokoľvek sa práve stalo, nezmenilo to tvoje vnútorné rozprávanie! Budeš naďalej hovoriť o vedomí z presne rovnakého dôvodu ako predtým.“

Bernice: „Títo malí roboti sú vládca zombií. Spôsobia, že budem hovoriť o vedomí, hoci nebudem mať vedomie. Svet zombií je možný, ak dovolíš pridanie navyše experimentálne zistiteľného vládcu zombií – a to *sú* tieto roboty.“

Charles: „Ach, to nie je pravda, Bernice. Tieto malé roboty neintrigujú ako napodobniť vedomie, ani nespracovávajú korpus textu od ľudských amatérov. Robia presne to isté, čo by robili neuróny, akurát pomocou kremíku namiesto uhlíku.“

Albert: „Moment, nehádal si sa so mnou pred chvíľou?“

Charles: „Nepovedal som, že by tá nová osoba nemala vedomie. Povedal som, že by som to nebol *ja*.“

Albert: „Nuž, antizombícky princíp vo všeobecnosti samozrejme hovorí, že táto operácia nenarušila skutočnú príčinu tvojho rozprávania o tvojom *ja*.“

Charles: „Och, och! Tvoja operácia určite narušila skutočnú príčinu môjho rozprávania o vedomí. Namiesto nej dosadila *inú* príčinu, tie roboty. Aj keď je táto nová príčina *tiež* vedomá – hovorí o vedomí z rovnakého *všeobecného* dôvodu – neznamená to, že je to *rovnaká* príčina ako bola pôvodne.“

Albert: „Ale ja by som ti o tejto robotickej operácii ani nemusel *povedať*. *Nevšimol* by si si to. Keď si myslíš, podľa introspektívnych indícií, že si v dôležitom zmysle slova ‚ten istý človek‘ ako si bol pred piatimi minútami, a ja urobím niečo, čo nezmení dostupné introspektívne indície, potom tvoj záver, že si ten istý človek ako pred piatimi minútami, by mal byť rovnako oprávnený. Nehovorí azda všeobecný antizombícky princíp, že ak ti urobím niečo, čo zmení tvoje vedomie, a tobôž ťa to urobí celkom odlišnou osobou, že by si si to mal nejako *všimnúť*?“

Bernice: „Nie ak ma nahradíš vládcom zombií. Potom tam nebude nikto, *kto* by si to všimol.“

Charles: „Introspekcia nie je dokonalá. V mozgu sa deje veľa vecí, ktoré si nevšímam.“

Albert: „Predpokladáte epifenomenálne fakty o vedomí a totožnosti!“

Bernice: „Ja nie! Môžem experimentálne zistiť rozdiel medzi neurónmi a robotmi.“

Charles: „Ja nie! Môžem experimentálne zistiť okamih, keď bolo moje staré ja nahradené novou osobou.“

Albert: „Áno, a ja dokážem zistiť, kedy bol prepnutý vypínač! Zisťujete niečo, čo *nerobí významný rozdiel v skutočnej príčine* vášho rozprávania o vedomí a osobnej totožnosti. A dôkazom je, že potom budete hovoriť celkom rovnako.“

Bernice: „Lebo to bude hovoriť tvoj robotický vládca zombií!“

Charles: „Aj keď dvaja ľudia hovoria o ‚osobnej totožnosti‘ z podobných dôvodov, nerobí to z nich tú istú osobu.“

Myslím si, že všeobecný antizombícky princíp podporuje Albertov postoj, ale dôvody budú musieť počkať na budúce články. Potrebujem ďalšie východiská a navyše je tento článok už dosť dlhý.

Ale vidíte dôležitosť otázky: „Ako ďaleko možno zovšeobecniť antizombícky princíp, aby stále platil?“

Od tejto odpovede môže závisieť zloženie budúcich galaktických civilizácií...

*→

224. VAZP verzus VVT

V článku „Nepredstaviteľná absurdita zombií“ Daniel Dennett hovorí:[205]

> Doteraz mi niekoľkí filozofi povedali, že plánujú prijať moju výzvu na poskytnutie obrany zombií, ktorá nepredpokladá svoj záver ako svoje východisko, ale jediná, ktorú som zatiaľ videl, zahŕňala predpoklad „logicky možnej“ ale fantastickej veci – potomka fantázie Neda Blocka o Veľkej Vyhľadávacej Tabuľke...

Veľká Vyhľadávacia Tabuľka, v reči programátorov, je keď naprogramujete funkciu ako veľkú tabuľku vstupov a výstupov, zvyčajne aby ste ušetrili na výpočtoch počas behu programu. Ak môj program potrebuje vedieť výsledok násobenia dvoch vstupov od 1 do 100, môžem napísať algoritmus násobenia, ktorý bude počítať vždy, keď sa táto funkcia zavolá, alebo môžem dopredu vypočítať Veľký Vyhľadávaciu Tabuľku s 10 000 vstupmi a dvoma indexmi. Sú situácie, keď *chcete* urobiť toto, aj keď nie pre násobenie – situácie, kde budete túto funkciu veľa používať a kde nemá veľa možných vstupov; alebo kde sú takty lacné pri štarte programu, ale veľmi drahé počas jeho behu.

Veľké Vyhľadávacie Tabuľky sa stanú veľmi veľkými veľmi rýchlo. VVT všetkých možných rozhovorov s 20 replikami, s 10 slovami na repliku, použijúc iba 850 slov základnej angličtiny, by si vyžadovala $7{,}6 \times 10^{585}$ vstupov.

Nahradiť ľudský mozog Veľkou Vyhľadávacou Tabuľkou všetkých možných zmyslových vstupov a pohybových výstupov (podľa nejakej jemnozrnnej digitalizačnej schémy) by si vyžadovalo *nehorázne veľké množstvo* pamäte. Ale „v princípe“, ako filozofi radi hovoria, by sa to dalo.

Táto VVT nie je zombia v klasickom zmysle, pretože je mikrofyzikálne nepodobná človeku. (V skutočnosti VVT nemôže *naozaj* fungovať podľa rovnakej fyziky ako človek; je taká veľká, že sa nezmestí do nášho vesmíru. Z filozofických dôvodov to budeme ignorovať a budeme predpokladať dostatočný pamäťový sklad.)

Ale je VVT *vôbec* zombia? To jest, správa sa presne ako človek, a pritom nemá vedomie?

Jazyk tela VVT hovorí o vedomí. Jeho prsty píšu filozofické články. V každom zmysle, dokiaľ sa nepozriete dovnútra jej lebky, sa VVT správa presne ako človek... čo iste vyzerá ako platný príklad zombie: správa sa ako človek, akurát nikto nie je doma.

Pokiaľ teda VVT nemá vedomie, lebo v tom prípade by to nebol platný príklad.

Nespomínam si, že by som videl *niekoho* tvrdiť, že VVT má vedomie. (Pripúšťam, že moje čítanie v tejto oblasti nie je na profesionálnej úrovni; nech sa páči, opravte ma.) Dokonca aj ľudia, ktorých obviňujú, že sú (ach!) funkcionalisti, netvrdia, že VVT má vedomie.

VVT sú *redukcia ad absurdum* pre každého, kto naznačí, že vedomie je *jednoducho* vzorom vstupov a výstupov, čím sa zbaví všetkých nepríjemných starostí ohľadom toho, čo sa deje vnútri.

Čo teda všeobecný antizombický princíp (VAZP) hovorí o veľkej vyhľadávacej tabuľke (VVT)?

Na prvý pohľad by sa zdalo, že VVT je archetypom vládcu zombií – samostatný, dodatočný, merateľný, nevedomý systém, ktorý oživuje zombiu a spôsobuje, že hovorí o vedomí z *iných* dôvodov.

Vo vnútri VVT je iba veľmi jednoduchý počítačový program, ktorý vyhľadáva vstupy a vracia výstupy. Dokonca aj hovoriť o „jednoduchom počítačovom programe“ v takomto prípade znamená preceňovať ho. VVT je skôr ako pamäť než ako procesor. Rovnako dobre by sme mohli hovoriť o postupnosti prepínačov na dráhe, po ktorej sa kotúľajú nejaké gule z pripraveného zásobníka do koryta – *bodka*; to je *všetko*, čo VVT robí.

Hovorca organizácie Ľudia za etické zaobchádzanie so zombiami odpovedá: „Ach, to je to, čo hovoria všetci antimechanisti, nie? Že keď sa pozriete na mozog, nájdete iba hŕstku neurotransmiterov

→ http://lesswrong.com/lw/p9/the_generalized_antizombie_principle/

205 Daniel C. Dennett, „The Unimagined Preposterousness of Zombies,“ *Journal of Consciousness Studies* 2 (4 1995): 322–26.

otvárajúcich iónové kanály? Ak iónový kanál môže mať vedomie, prečo nie páky a gule kotúľajúce sa do nádob?“

„Problém nie sú tie páky,“ odpovie funkcionalista, „problém je, že VVT má *nesprávny vzor* pák. Potrebujete páky, ktoré implementujú veci ako napríklad vytváranie názorov o názoroch, alebo sebamodelovanie... Do kelu, potrebujete schopnosť zapisovať veci do pamäte, aby mohol vôbec plynúť čas výpočtu. Pokiaľ si nemyslíte, že je možné naprogramovať vedomú bytosť v Haskelli.“

„O tom nič neviem,“ povie hovorca ĽZEZSZ, „viem iba, že táto takzvaná zombia píše filozofické články o vedomí. Odkiaľ pochádzajú tieto filozofické články, ak nie z vedomia?“

Dobrá otázka! Zamyslime sa nad ňou hlbšie.

Vo fyzike existuje hra nazývaná Sleduj energiu. Hrával ju otec Richarda Feynmana s malých Richardom:

> Bola to jedna z tých vecí, o ktorých zvykol hovoriť môj otec: „Prečo sa to hýbe? Všetko sa hýbe preto, lebo svieti slnko.“ A potom sme sa bavili rozhovorom o tom:
>
> „Nie, hračka funguje, lebo má natiahnutú pružinu,“ povedal by som. „Ako sa pružina natiahla?“ opýtal by sa.
>
> „Ja som ju natiahol.“
>
> „A prečo sa ty hýbeš?“
>
> „Lebo jem.“
>
> „A jedlo rastie iba preto, lebo slnko svieti. Takže všetky teto veci sa hýbu preto, lebo slnko svieti.“ Toto by vysvetlilo predstavu, že pohyb je jednoducho *transformácia* slnečnej energie.[206]

Keď trochu vyrastiete, dozviete sa, že energia sa zachováva, nikdy sa nevytvára ani neničí, takže pojem *spotreby* energie nedáva veľmi zmysel. Nikdy nemôžete zmeniť celkové množstvo energie, tak v akom zmysle ju *spotrebovávate*?

Takže keď fyzici vyrastú, naučia sa hrať novú hru s názvom Sleduj negentropiu – čo je v skutočnosti tá istá hra, ktorú hrali doteraz; iba jej pravidlá sú matematickejšie, hra je užitočnejšia, a jej princípy je ťažšie pochopiť.

Racionalisti sa naučia hru s názvom Sleduj nepravdepodobnosť, čo je verzia hry „Odkiaľ to vieš?“ pre dospelých. Pravidlo racionalistickej hry je, že každý nepravdepodobne vyzerajúci názor potrebuje ekvivalentné množstvo indície na jeho zdôvodnenie. (Táto hra má *úžasne podobné* pravidlá ako Sleduj negentropiu.)

Kedykoľvek niekto poruší pravidlá racionalistickej hry, môžete nájsť v jeho argumente miesto, kde sa množstvo nepravdepodobnosti objavuje odnikiaľ; a to je rovnakým príznakom problému ako povedzme zložitý dizajn spojených kolies a ozubených kolies, ktoré sa sami od seba večne otáčajú.

Niekto príde za vami a povie: „Verím pevnou a trvalou vierou, že v pásme asteroidov je predmet, jednu stopu veľký a celý zložený z čokoládových koláčov; nemôžeš mi dokázať, že je to nemožné.“ Lenže, pokiaľ dotyčný nemá prístup k nejakému druhu indície pre tento názor, bolo by vysoko nepravdepodobné, že by sa *spontánne* vytvoril presný názor. Takže buď dotyčný ukáže na indície, alebo sa tento názor neukáže ako pravdivý. „Ale ty nemôžeš dokázať, že je *nemožné*, aby si moja myseľ spontánne vytvorila názor, ktorý bude zhodou okolností pravdivý!“ Nie, ale takéto spontánne vygenerovanie je *vysoko nepravdepodobné*, asi ako povedzme rozbité vajce, ktoré sa samo poskladá.

V hre Sleduj nepravdepodobnosť je vysoko podozrivé čo len *hovoriť* o konkrétnej hypotéze bez dostatočnej indície na zúženie priestoru možných hypotéz. Prečo neposkytujete rovnaký mediálny priestor deciliónu iných rovnako pravdepodobných hypotéz. Potrebujete dostatok indície, aby ste našli hypotézu „čokoládový koláč v pásme asteroidov“ v priestore hypotéz – inak by neexistoval dôvod, prečo

206 Richard P. Feynman, „Judging Books by Their Covers,“ in *Surely You're Joking, Mr. Feynman!* (New York: W. W. Norton & Company, 1985).

jej dať viac mediálneho priestoru než triliónom iných kandidátov ako „V pásme asteroidov je drevená skrinka“ alebo „Lietajúce špagetové monštrum sa vyvracalo na moje tenisky.“

V hre Sleduj nepravdepodobnosť nemáte dovolené vyťahovať veľké zložité konkrétne hypotézy z klobúka, pokiaľ *už* nemáte zodpovedajúce množstvo indície; pretože nie je realistické predpokladať, že by ste mohli spontánne začať diskutovať o *pravdivej* hypotéze čírou *zhodou okolností*.

Filozof povie: „Lebka zombie obsahuje Veľkú Vyhľadávaciu Tabuľku všetkých vstupov a výstupov ľudského mozgu.“ Toto je veľmi *veľká* nepravdepodobnosť. Opýtate sa teda: „Ako sa stala táto nepravdepodobná udalosť? Odkiaľ prišla táto VVT?“

Toto veru nie je štandardný filozofický postup pri myšlienkových experimentoch. Pri štandardnom filozofickom postupe máte povolené tvrdiť veci ako „Predstav si, že cestuješ na lúči svetla...“ bez obáv o fyzikálnu možnosť, a tobôž púhu pravdepodobnosť. Ale v tomto prípade na pôvode VVT záleží; a preto je dôležité pochopiť motivujúcu otázku: „Odkiaľ prišla táto nepravdepodobnosť?“

Samozrejmá odpoveď je, že ste si vzali komplikovanú špecifikáciu ľudského mozgu, a *tú* ste použili na predpočítanie VVT. (Čím ste vytvorili nespočetné množstvá ľuských bytostí, niektoré v extrémnych bolestiach, drvivú väčšinu pomerne šialenú vo vesmíre chaosu, kde vstupy nijako nesúvisia s výstupmi. Ale kašľať na etiku, toto je v záujme *filozofie*.)

V tomto prípade VVT *píše* články o vedomí kvôli vedomému algoritmu. VVT nie je zombia o nič viac než mobilný telefón, ktorý tiež dokáže rozprávať o vedomí, hoci je to iba malé spotrebné elektronické zariadenie. Mobilný telefón iba prenáša filozofické prejavy od toho, kto je na druhej strane linky. VVT pôvodne vytvorená podľa špecifikácie ľudského mozgu robí to isté.

„V poriadku,“ povie filozof, „VVT bola vytvorená náhodne, a *iba zhodou okolností* má rovnaké vzťahy medzi vstupmi a výstupmi ako nejaký referenčný človek.“

Ako presne ste náhodne vygenerovali túto VVT?

Použili sme zdroj skutočnej náhody – kvantové zariadenie.

Lenže kvantové zariadenie iba implementuje príkaz Choď Obidvoma Cestami; keď vytvoríte bit zo zdroja kvantovej náhodnosti, deterministický výsledok je, že jedna množina vetiev vesmíru (lokálne spojených oblakov amplitúdy) vidí 1, a druhá množina vesmírov vidí 0. Urobte to 4-krát, vytvoríte 16 (množín) vesmírov.

Takže v skutočnosti je to ako povedať, že ste vytvorili túto VVT tak, že ste vypísali všetky možné postupnosti 0 a 1 veľkosti VVT, čo bola naozaj sakra veľká hora vyhľadávacích tabuliek; a potom ste siahli do tejto hory a *nejako* ste vybrali VVT, ktorá zhodou okolností zodpovedá špecifikácii ľudského mozgu. Odkiaľ prišla táto nepravdepodobnosť?

Pretože ak toto *nebola iba zhoda okolností* – ak ste mali nejakú funkciu siahni-do-hory, ktorá vytiahla VVT zodpovedajúcu človeku podľa svojej *funkcionality*, nie iba náhodou – potom je táto funkcia siahni-do-hory pravdepodobne vedomá, takže VVT je opäť ako mobil, nie ako zombia. Je spojená s človekom na dva kroky, nie na jeden, ale stále je to telefón! Pekný pokus o skrytie zdroja nepravdepodobnosti!

Pozrime teraz, kam nás zaviedlo Sleduj nepravdepodobnosť: kde je ten príčina, prečo jazyk tohto tela hovorí o vnútornom poslucháčovi? Vedomie nie je vo vyhľadávacej tabuľke. Vedomie nie je v továrni, ktorá vyrába množstvo možných vyhľadávacích tabuliek. Vedomie je v tom, čo *ukázalo na jednu konkrétnu spomedzi už vyrobených vyhľadávacích tabuliek* a povedalo: „Použi *túto*!“

Vidíte, prečo som zaviedol hru Sleduj nepravdepodobnosť. Keď sa za bežných okolností s niekým rozprávame, máme sklon si myslieť, že nech je v jeho lebke čokoľvek, musí to byť „to, kde je vedomie“. Iba hraním Sleduj pravdepodobnosť si uvedomíme, že skutočný zdroj konverzácie, ktorú máme, je to-čo-je-zodpovedné-za *nepravdepodobnosť* v konverzácii – nech je to akokoľvek vzdialené v čase alebo priestore, ako keď Slnko poháňa hračku s pružinou.

„Nie, nie!“ povie filozof. „V tomto myšlienkovom experimente nevytvárame náhodne hromady VVT a nepoužívame potom vedomý algoritmus, aby vybral jednu VVT, ktorá vyzerá ľudsky! Ja som *špecifikoval*, že v tomto myšlienkovom experimente siahnu do nepredstaviteľne rozsiahlej hory VVT a

čírou zhodou okolností vyriahnu VVT, ktorá je totožná so vstupmi a výstupmi ľudského mozgu! *Tak!* Teraz som ťa dostal! Už nemôžeš ďalej hrať Sleduj nepravdepodobnosť!“

Aha, takže tvoja *špecifikácia* je tu zdrojom nepravdepodobnosti.

Keď sa opäť zahráme Sleduj nepravdepodobnosť, skončíme *mimo myšlienkového experimentu*, pozerajúc na samotného *filozofa*.

To, čo ukazuje na tú jednu VVT, ktoré hovorí o vedomí, spomedzi celého rozľahlého priestoru možností, je teraz... tá vedomá osoba, ktorá nás žiada, aby sme si predstavili celý tento scenár. A náš vlastný mozog, ktorý dopĺňa prázdne miesta, keď si predstavujeme: „Čo by táto VVT povedala v reakcii na: ‚Povedz mi o svojom vnútornom poslucháčovi.‘?“

Ponaučenie z tohto príbehu je, že keď sledujete naspäť debatu o „vedomí“, vo všeobecnosti nájdete vedomie. Nie vždy je priamo pred vami. Niekedy je veľmi šikovne ukryté. Ale je tam. Preto všeobecný antizombícky princíp.

Ak existuje vládca zombií vo forme diskusného robota, ktorý spracováva a remixuje debaty amatérskych ľudí o „vedomí“, potom sú vedomí tí ľudia, ktorý vytvorili pôvodný korpus textu.

Ak jedného dňa pochopíte, čo je to vedomie, a obzriete sa a uvidíte, že existuje program, ktorý dokážete napísať, ktorý bude vypisovať zmätené filozofické debaty, ktoré budú znieť celkom ako ľudia, hoci sám nebude vedomý – potom keď sa opýtam: „Ako je možné, že tento program znie tak podobne ľuďom?“, odpoveď je, že to *vy* ste ho napísali tak, aby znel podobne *vedomým ľuďom*, namiesto aby ste si vybrali kritérium podobnosti s niečím iným. Neznamená to, že váš malý vládca zombií má vedomie – ale znamená to, že niekde vo vesmíre nájdeme vedomie stopovaním späť podľa reťaze kauzality, čo znamená, že sa celkom nenachádzame vo svete zombií.

Ale čo keby niekto naozaj *siahol* do hory VVT a *skutočnou čírou náhodou* by vytiahol VVT, ktorá píše filozofické články?

Nuž, potom by nebola vedomá. Podľa môjho skromného názoru.

Myslím tým, že v tom musí byť viac než iba vstupy a výstupy.

Inak by aj VVT mohla byť vedomá, však?

Aha, a pre tých, ktorí sa čudujú, ako sa toto celé týka mojej každodennej práce...

V tomto biznise stretnete strašne veľa ľudí, ktorí si myslia, že náhodne vygenerovaná UI bude „morálna“. Nedokážu sa zhodnúť na tom, prečo, alebo čo myslia slovom „morálna“; ale všetci sa zhodnú na tom, že robiť Priateľskú UI je zbytočné. A keď sa ich opýtate, ako náhodne vytvorená UI skončí s morálnymi výstupmi, ponúknu vám prepracované racionalizácie určené pre UI ohľadom toho, čo oni považujú za „morálne“; a v tom sú všelijaké druhy problémov, ale problém číslo jeden je: „Ste si *istí*, že UI by sa riadila rovnakou líniou myslenia, akú ste si vymysleli na zdôvodnenie ľudskej morálky, keď na rozdiel od vás daná UI nezačne s vedomosťou, *čo* chcete, aby si racionalizovala?“ Mohli by ste tento protiprincíp nazvať Sleduj-Informáciu-O-Rozhodnutí alebo také niečo. Môžete mi rozprávať o UI, ktorá robí nepravdepodobne pekné veci, ak mi poviete, ako vyberiete dizajn tejto UI z obrovského priestoru možností, pretože inak vyťahujete nepravdepodobnosť z klobúka – hoci je lepšie a lepšie maskovaná, ako si racionalizujete východiská ďalších racionalizácií.

Tak už som napísal celú sériu článkov, ktoré som sám vygeneroval pomocou hry Sleduj nepravdepodobnosť. Ale vtedy som ešte tieto pravidlá nevypísal *explicitne*, pretože som ešte neurobil články o termodynamike...

Len som sa rozhodol to spomenúť. Je úžasné, pri koľkých spomedzi mojich článkov sa zhodou okolností ukáže, že obsahujú myšlienky týkajúce sa diskusie o Priateľskej UI... ak veríte na zhody okolností.

*→

→ http://lesswrong.com/lw/pa/gazp_vs_glut/

225. Viera v implicitné neviditeľné

Jedna všeobecná lekcia, ktorý by ste sa *nemali* naučiť z argumentu proti zombiám je: „Keď to nemôžete vidieť, tak to neexistuje.“

Je vábivé prijať toto všeobecné pravidlo. Urobilo by to v budúcnosti argument proti zombiám omnoho jednoduchším, keby sme to mohli použiť ako predpoklad. Ale žiaľ, to skrátka nie je bayesovské.

Predstavme si, že vyšlem fotón smerom do nekonečna, nenamierim ho na žiadnu hviezdu ani galaxiu, namierom ho smerom k jednej z veľkých prázdnot medzi superzhlukmi. Inými slovami, podľa štandardnej fyziky nepredpokladám, že sa tento fotón cestou preč s niečím zrazí. Fotón sa hýbe rýchlosťou svetla, takže ho nemôžem dobehnúť a opäť chytiť.

Ak sa rozširovanie vesmíru zrýchľuje, ako tvrdí súčasná kozmológia, v budúcnosti nastane bod, v ktorom nebude očakávať, že by som dokázal s týmto fotónom interagovať hoci len v princípe – budúcnosť, v ktorej neočakávam, že sa svetelný kužeľ tohto fotónu pretne s mojou svetočiarou. Dokonca aj keby mimozemšťania zachytili tento fotón a poslali ho späť k nám, necestoval by dosť rýchlo na to, aby prekonal zrýchľujúce sa rozširovanie vesmíru.

Mal by som teda veriť, že v okamihu, keď už s ním viac nemôžem interagovať ani v princípe, fotón zmizol?

Nie.

Porušilo by to zákon zachovania energie. A druhý zákon termodynamiky. A prakticky každý ďalší fyzikálny zákon. Možno aj tri zákony robotiky. Naznačovalo by to, že fotón vie, že sa oň zaujímam, a vie presne, kedy má zmiznúť.

To je *hlúposť*.

Ale ak môžete veriť v pokračujúcu existenciu fotónu, ktorý sa pre vás stal experimentálne nezistiteľným, prečo z toho nevyplýva všeobecné právo veriť v neviditeľné veci?

(Ak sa nad touto otázkou chcete zamyslieť sami, urobte tak pred nasledujúcicm čítaním...)

Hoci sa mi nepodarilo vygoogliť zdroj, pamätám sa, ako som čítal, keď sa prvýkrát uvažovalo, že Mliečna dráha je naša *galaxia* – že tá hmlistá rieka svetla na nočnej oblohe sa skladá z miliónov (alebo dokonca miliárd) hviezd – že proti tejto novej hypotéze použili Occamovu britvu. Pretože, ako vidíte, táto hypotéza značne znásobila počet „entít“ v domnelom vesmíre. Alebo možno to bola úvaha, že „hmloviny“ - tie hmlisté škvrny viditeľné ďalekohľadom – by mohli byť galaxie plné hviezd, na ktorú použili Occamovu britvu.

Lex parsimoniae: Entia non sunt multiplicanda praeter necessitatem.

Toto bola pôvodná Occamova formulácia, zákon šetrnosti: Veci by sa nemali množiť viac než je nutné.

Ak predpokladáte miliardy hviezd, v ktoré pred vami nikto neveril, množíte tým veci, nie?

Nie. Existujú dve bayesovské formulácie Occamovej britvy: Solomonoffova indukcia a Minimálna dĺžka správy. Žiadna z nich nepenalizuje galaxie za to, že sú veľké.

Čo by veru nemali robiť! Jedna z lekcií dejín je, že to-čo-voláme-skutočnosť sa stále ukazuje ako väčšie a väčšie a ešte obrovskejšie. Pamätáte sa, keď bola Zem stredom vesmíru? Pamätáte sa, keď nikto nepoznal Avogadrovo číslo? Keby Occamova britva zakaždým pôsobila proti množeniu vecí, museli by sme o nej začať pochybovať, lebo by sa konzistentne ukazovala ako nesprávna.

V Solomonoffovej indukcii je zložitosť vášho modelu množstvo *kódu* v počítačovom programe, ktorý by ste museli napísať, aby simuloval váš model. Množstvo *kódu*, nie množstvo spotrebovanej pamäte, ale počet taktov, ktoré vyžaduje na výpočet. Model vesmíru, ktorý obsahuje miliardy galaxií obsahujúcich miliardy hviezd, kde sa každá hviezda skladá z biliónu triliónov kvadriliónov kvarkov, spotrebuje veľa pamäte – ale samotný *kód* musí iba opisovať správanie kvarkov, a hviezdy a galaxie už môžeme nechať, nech bežia samé. Hovorím to napoly metaforicky – vo vesmíre existujú aj iné veci ako kvarky – ale pointa je, že predpoklad miliardy galaxií navyše sa nepočíta proti dĺžke vášho kódu, ak už ste opísali jednu galaxiu. Iba to zaberie viac pamäte a Occamova britva sa o pamäť nestará.

Prečo nie? Formalizmus Miminálnej dĺžky správy, ktorý je takmer ekvivalentný Solomonoffovej indukcii, ten princíp možno objasní lepšie: Ak musíte niekomu povedať, ako funguje váš model vesmíru, nemusíte jednotlivo zadať polohu každého kvarku v každej hviezde v každej galaxii. Iba napíšete nejaké rovnice. Množstvo „hmoty“, ktoré sa riadi touto rovnicou, neovplyvňuje ako dlho vám bude trvať túto rovnicu napísať. Ak zakódujete túto rovnicu do súboru a súbor má dĺžku 100 butov, potom existuje 2^{100} iných modelov, ktorých súbor by mal zhruba rovnakú dĺžku a vy potrebujete zhruba 100 bitov podporujúcej indície. Máte obmedzené množstvo masy pravdepodobnosti; apriori musíte túto masu rozdeliť medzi všetky správy, ktoré môžete poslať; takže predpoklad modelu z priestoru modelov s 2^{100} alternatívami znamená, že musíte prijať apriórnu penaltu pravdepodobnosti 2^{-100} – ale mať viac galaxií k tomu nepridáva.

Predpoklad miliárd hviezd v miliardách galaxií neovplyvňuje dĺžku vaše správy popisujúcej celkové správanie týchto galaxií. Nedostanete teda pravdepodobnostný zásah za to, že *tie isté* rovnice opisujú viac vecí. (Pokiaľ úspešnosť predpovedania vášho modelu nie je citlivá na presné počiatočné podmienky. Ak vo vašom modeli musíte zadať presné polohy všetkých kvarkov, aby predpovedal tak dobre, ako to teraz robí, tieto kvarky navyše sa počítajú ako zásah.)

Ak predpokladáte, že fotón zmizne, keď sa naň prestanete pozerať, toto je *dodatočný zákon* vo vašom modeli vesmíru. Zákony sú tie „veci“, ktoré zákon šetrnosti považuje za drahé. Kvarky navyše sú zadarmo.

Takže sa to celé zúžilo na: „Verím tomu, že fotón pokračuje v existencii ako odlieta do prázdnoty, pretože moje pôvodné pravdepodobnosti hovoria, že je jednoduchšie, ak bude existovať, než keby zmizol“?

Tak som si na prvý pohľad myslel, ale po zamyslení to nie je celkom správne. (A nie iba preto, lebo to otvára dvere zrejmému zneužívaniu.)

Zúžil by som to na rozdiel medzi vierou v *implicitné neviditeľné* a vierou v *dodatočné neviditeľné*.

Keď veríte, že fotón pokračuje v existencii aj keď odlieta do nekonečna, neveríte v to ako v *dodatočný* fakt.

To, v čo veríte (priraďujete pravdepodobnosť) je množina jednoduchých rovníc; veríte, že tieto rovnice opisujú vesmír. Veríte v tieto rovnice, pretože sú to najjednoduchšie rovnice, ktoré ste našli, ktoré opisujú dané indície. Tieto rovnice sú *vysoko* experimentálne testovateľné; vysvetľujú obrovské hromady indícií viditeľných v minulosti, a predpovedajú výsledky mnohých pozorovaní v budúcnosti.

Veríte v tieto rovnice, a je *logickým dôsledkom* týchto rovníc, že daný fotón pokračuje v existencii ako odlieta do prázdnoty, takže veríte aj tomu.

Vaše pôvodné pravdepodobnosti, dokonca ani vaše terajšie pravdepodobnosti nehovoria *priamo* o tomto fotóne. To, čomu priraďujete pravdepodobnosť, nie je daný fotón, ale všeobecné zákony. Keď priradíte pravdepodobnosť zákonom fyziky, ako ich poznáme, *automaticky* pripisujete rovnakú pravdepodobnosť tomu, že fotón pokračuje v existencii ako odlieta do prázdnoty – ak veríte v logické dôsledky toho, v čo veríte.

Nie je to tak, že veríte v neviditeľné *ako také*, vďaka uvažovaniu o neviditeľných veciach. Skôr experimentálne indície podporujú určité zákony, a z viery v tieto zákony logicky vyplýva existencia určitých vecí, s ktorými nemôžete interagovať. Toto je viera v *implicitné neviditeľné*.

Na druhej strane, ak veríte tomu, že tento fotón pohltí z existencie lietajúce špagetové monštrum – možno iba v tomto jedinom prípade – ale dokonca ak bezdôvodne veríte, že tento fotón cestou preč narazí na zrnko prachu – potom by ste osobitne verili v nejakú konkrétnu neviditeľnú udalosť navyše. Keby ste si mysleli, že sa takéto veci dejú všeobecne, potom by ste verili v nejaký konkrétny neviditeľný zákon navyše. Toto je viera v *dodatočné neviditeľné*.

Aby bolo jasnejšie, prečo niekedy potrebujete myslieť na implicitné neviditeľné, predstavte si, že idete vyslať vesmírnu loď rýchlosťou blízkej rýchlosti svetla, smerom k vzdialenému superzhluku. V čase, keď sa tam vesmírna loď dostane a založí tam kolóniu, bude rozpínanie vesmíru natoľko zrýchlené, že nemôžu poslať správu naspäť. Myslíte si, že má zmysel čisto altruistické úsilie vytvoriť túto kolóniu,

kvôli všetkým tým ľuďom, ktorí tam budú šťastne žiť? Alebo si myslíte, že vesmírna loď zmizne z existencie skôr než sa tam dostane? Toto môže byť jedného dňa celkom skutočná otázka.

Pripúšťam, že celá vec by bola omnoho jednoduchšia, keby sme mohli jednoducho vylúčiť existenciu vecí, s ktorými nemôžeme interagovať, raz a navždy – aby vesmír končil na hranici našich ďalekohľadov. Ale to by sme museli byť veľmi pochabí.

Povedať, že by ste nemali nikdy potrebovať samostatnú a dodatočnú vieru v neviditeľné veci – že veríte iba v neviditeľné veci, ktoré sú *logickými dôsledkami* všeobecných zákonov, ktoré samotné sú testovateľné a že ešte aj vtedy nemáte žiadne ďalšie názory na ne, ktoré nie sú logickými dôsledkami viditeľne testovateľných všeobecných pravidiel – toto zdá sa naozaj vylučuje všelijaké zneužívanie viery v neviditeľné, ak to použijete správne.

Možno by som mal povedať: „mali by ste dodatočnému neviditeľnému priradiť nezmenenú pôvodnú pravdepodobnosť“ namiesto: „neverte v to“. Ale ak myslíte na *vieru* ako na niečo dodatočné z hľadiska indícií, niečo, čo sa unúvate sledovať, niečo, kde sa unúvate počítať podporu za alebo proti, potom je otázne, či by sme vôbec niekedy mali dodatočne veriť na dodatočné neviditeľné.

Existujú exotické prípady, ktoré túto teóriu porušujú. (Napríklad: Epifenomenálni démoni vás pozorujú a budú mučiť 3^^^3 obetí celý rok, niekde, kde si túto udalosť nemôžete uveriť, ak niekedy vyslovíte slovo „Niblick“.) Ale nenapadá mi prípad, kde by tento princíp zlyhával v ľudskej praxi.

*→

—

226. Zombie: Film

ZAOSTRENIE na vážne vyzerajúcu skupinu vojenských dôstojníkov v uniformách. Na čele stola hovorí starší zavalitý muž, GENERÁL FRED.

GENERÁL FRED: Správy sa potvrdili. Ney York zaplavili... *zombie*.

PLUKOVNÍK TODD: Opäť? Veď sme mali inváziu zombií iba pred 28 dňami!

GENERÁL FRED: Tieto zombie... sú iné. Sú to.... *filozofické* zombie.

KAPITÁN MUDD: Sú plné hnevu, ktorý ich núti hrýzť ľudí?

PLUKOVNÍK TODD: Stratili všetku schopnosť rozumu?

GENERÁL FRED: Nie. Správajú sa... *celkom* rovnako ako my... akurát, že nemajú vedomie.

(Pri stole zavládne ticho.)

PLUKOVNÍK TODD: Dobrý bože.

GENERÁL FRED prejde k displeju s počítačom.

GENERÁL FRED: Toto je mesto New York, pred dvoma týždňami.

Displej ukazuje davy hemžiace sa ulicami, ľudí jediacich v reštauráciách, smetiarske vozidlo odťahujúce smeti.

GENERÁL FRED: *Toto*... je mesto New York... *teraz*.

Displej sa zmení, ukáže preplnený vlak metra, skupinu študentov smejúcich sa v parku, a dvojicu držiaci sa za ruky v slnečnom svetle.

PLUKOVNÍK TODD: Je to horšie, než som si predstavoval.

KAPITÁN MUDD: Ako presne to dokážete rozoznať?

PLUKOVNÍK TODD: Nikdy som nevidel nič tak brutálne všedné.

VEDEC v laboratórnom plášti vstane na opačnom konci stola.

VEDEC: Choroba zombií odstráni vedomie bez toho, že by ľubovoľným spôsobom zmenila mozog. Pokúšali sme sa pochopiť, ako sa táto choroba prenáša. Došli sme k záveru, že keďže táto choroba napáda duálne vlastnosti obyčajnej hmoty, musí samotná fungovať mimo nášho vesmíru. Máme tu do činenia s *epifenomenálnym vírusom*.

GENERÁL FRED: Ste si tým istí?

VEDEC: Takí istí, ako si len môžeme byť pri naprostom nedostatku indícií.

→ http://lesswrong.com/lw/pb/belief_in_the_implied_invisible/

GENERÁL FRED: V poriadku. Zostavte hlásenie o každom epifenoméne, ktorý bol kedy pozorovaný. Čo, kde, a kto. Chcem zoznam všetkého, čo sa nestalo počas posledných päťdesiatich rokov.

KAPITÁN MUDD: Ak je ten vírus epifenomenálny, ako vieme, že existuje?

VEDEC: Rovnako ako vieme, že *máme* vedomie.

KAPITÁN MUDD: Aha, dobre.

GENERÁL FRED: Dosiahli lekári nejaký pokrok v hľadaní epifenomenálneho lieku?

VEDEC: Vyskúšali každé placebo v knihách. Nevyšlo. Všetko, čo robia, má nejaký účinok.

GENERÁL FRED: Našli ste nejakého homeopata?

VEDEC: Skúšal som pane! Nemohol som žiadneho nájsť!

GENERÁL FRED: Výborne. A čo taoisti?

VEDEC: Odmietajú čokoľvek urobiť!

GENERÁL FRED: Potom ešte máme šancu na záchranu.

PLUKOVNÍK TODD: A čo David Chalmers? Nemal by tu byť?

GENERÁL FRED: Chalmers... bol jednou z prvých obetí.

PLUKOVNÍK TODD: Ach nie.

(*Strih* na INTERIÉR kobky, celej ohraničenej vystuženým sklom, kde sa DAVID CHALMERS prechádza tam a späť.)

LEKÁR: David! David Chalmers! Počujete ma?

CHALMERS: Áno.

SESTRA: Je to zbytočné, pán doktor.

CHALMERS: Som celkom v poriadku. Skúmal som introspekciou svoje vedomie, a neviem nájsť žiaden rozdiel. *Viem*, že by som očakával, že to poviem, ale...

LEKÁR sa s hrôzou odvráti od sklenenej obrazovky.

LEKÁR: Jeho slová... *neznamenajú nič.*

CHALMERS: Toto je groteskné skreslenie mojich filozofických názorov. Takáto vec sa nemôže naozaj stať!

LEKÁR: Prečo nie?

SESTRA: Áno, prečo nie?

CHALMERS: Pretože...

(*Strih* na dvoch POLICAJTOV strážiacich prašnú cestu vedúcu k impozantnej oceľovej bráne gigantického betónového komplexu. Visačka na ich uniformách uvádza: „AGENTÚRA NA UPLATŇOVANIE PREMOSŤOVACÍCH ZÁKONOV“.)

POLICAJT 1: Musíš si dávať bacha na týchto chytrých bastardov. Vyzerajú ako ľudia. Môžu hovoriť ako ľudia. Sú s ľuďmi totožní na atómovej úrovni. Ale nie sú to ľudia.

POLICAJT 2: Darebáci.

Veľký hluk burácajúceho motora sa ozýva ponad kopce. Prichádza MUŽ na bielej motorke. MUŽ má oblečené čierne okuliare, čierny kožený oblek s čiernou koženou kravatou, a strieborné kovové topánky. Biela brada mu veje vo vetre. Zastaví pred bránou.

POLICAJTI pribehnú k motorke.

POLICAJT 1: Uveďte dôvod vašej prítomnosti.

MUŽ: Tu držíte Davida Chalmersa?

POLICAJT 2: Čo je vás do toho? Ste jeho priateľ?

MUŽ: Nemôžem povedať, že by som bol. Ale aj zombie majú práva.

POLICAJT 1: Dobre, kamoš, chceme vidieť tvoje kvaliá.

MUŽ: Žiadne nemám.

POLICAJT 2 náhle vytiahne pištoľ a drží ju namierenú na MUŽA.

POLICAJT 2: Aha! Zombia!

POLICAJT 1: Nie, zombie tvrdia, že majú kvaliá.

POLICAJT 2: Takže je to obyčajný človek?

POLICAJT 1: Nie, tí tiež tvrdia, že majú kvaliá.

POLICAJTI pozerajú na MUŽA, ktorý pokojne čaká.

POLICAJT 2: Ehm...

POLICAJT 1: Kto *ste*?

MUŽ: Som Daniel Dennett, somári.

Zdanlivo odnikiaľ DENNETT vytiahne meč a s oceľovým zvukom pretne napoly pištoľ POLICAJTA 2. POLICAJT 1 siahne po svojej pištoli, ale DENNETT náhle stojí za POLICAJTOM 1, udrie päsťou a zasiahne POLICAJTA 1 na miesto medzi plecom a krkom. POLICAJT 1 padne na zem.

POLICAJT 2 cúva, zhrozený.

POLICAJT 2: To nie je možné! Ako ste to urobili?

DENNETT: Som totožný so svojím telom.

DENNETT ďalším úderom zloží POLICAJTA 2 a kráča smerom k bráne. Pozrie hore na impozantný betónový komplex a silnejšie stisne svoj meč.

DENNETT *(potichu, sám pre seba)*: Lyžica existuje.

(*Strih* naspäť na GENERÁLA FREDA a ostatných vojenských dôstojníkov.)

GENERÁL FRED: Práve som dostal hlásenie. Prišli sme o Detroit.

KAPITÁN MUDD: Nechcem byť ten, ktorý povie „Dobre sa stalo“, ale...

GENERÁL FRED: Austrália bola... *zredukovaná na atómy*.

PLUKOVNÍK TODD: Ten epifenomenálny vírus sa šíri rýchlejšie. Hrozí, že sa samotná civilizácia rozpustí do úplnej normálnosti. Možno sa práve pozeráme na stred ľudstva.

KAPITÁN MUDD: Môžeme so zombiami nejako vyjednávať?

GENERÁL FRED: Poslali sme im správy. Oni nám poslali iba jednu odpoveď.

KAPITÁN MUDD: A bolo to...?

GENERÁL FRED: Je práve na ceste.

Posol prinesie obálku a odovzdá ju GENERÁLOVI FREDOVI.

GENERÁL FRED otvorí obálku, vyberie jediný list papiera a číta ho.

V miestnosti zavládne ticho.

KAPITÁN MUDD: Čo hovoria?

GENERÁL FRED: Hovoria... že to *my* sme tí, ktorí majú vírus.

(Zavládne ticho.)

PLUKOVNÍK TODD zdvihne ruky a pozerá na ne.

PLUKOVNÍK TODD: Bože môj, je to pravda. Je to pravda. Ja...

(Po líci PLUKOVNÍKA TODDA steká slza.)

PLUKOVNÍK TODD: Ja nič necítim.

Obrazovka stmavne.

Zvuk stíchne.

Film pokračuje úplne rovnako ako dovtedy.

*→

—

227. Vylúčenie nadprirodzena

Z času na čas počujete niekoho tvrdiť, že by sa kreacionizmus nemal vyučovať v školách, najmä nie ako konkurenčná hypotéza k evolúcii, pretože kreacionizmus je *a priori a automaticky* vylúčený z vedeckého hľadiska, pretože sa odvoláva na „nadprirodzeno“.

Takže... znamená to, že kreacionizmus by *mohol* byť pravdivý, ale *aj keby bol pravdivý*, nebolo by *dovolené* učiť ho na hodinách prírodovedy, pretože veda je iba o „prirodzených“ veciach?

Zdá sa dosť jasné, že táto predstava vychádza z túžby vyhnúť sa konfrontácii medzi vedou a náboženstvom. Nechcete na rovinu prísť a povedať, že veda neučí Náboženské Tvrdenie X, pretože X bolo testované vedeckou metódou a ukázalo sa ako nepravdivé. Namiesto toho teda môžete... ehm...

→ http://lesswrong.com/lw/pn/zombies_the_movie/

tvrdiť, že veda vylučuje hypotézu X a priori. Tak sa vyhnete diskusii o tom, ako experiment falzifikoval X a posteriori.

To samozrejme nahráva do kariet kreacionistického tvrdeniu, že Inteligentný Dizajn nedostáva spravodlivý priestor v prírodovede – že veda má v tejto veci *predsudky* v prospech ateizmu, bez ohľadu na indície. Keby veda naozaj vylučovala Inteligentný Dizajn a priori, bola by to oprávnená výhrada!

Ale vráťme sa o kúsok. Niekto k vám príde a povie: „Inteligentný Dizajn je a priori vylúčený z vedy, pretože je ‚nadprirodzený' a veda pracuje iba s ‚prirodzenými' vysvetleniami."

Čo presne myslia slovom „nadprirodzený"? Je ľubovoľné vysvetlenie, ktoré vymyslí niekto s priezviskom „Cohen" nadprirodzené? Ak chceme súhrnne vykopnúť z vedy nejakú množinu hypotéz, čo presne je to, čo ideme vylúčiť?

Zatiaľ *zďaleka* najlepšia definícia nadprirodzena, akú som kedy počul, je od Richarda Carriera: „Nadprirodzené" vysvetlenie sa odvoláva na *ontologicky základné myšlienkové veci*, myšlienkové veci, ktoré nemožno redukovať na nemyšlienkové veci.

Toto je rozdiel, napríklad, medzi tvrdením, že voda tečie dole kopcom preto, lebo *chce* byť nižšie, a predložením diferenciálnych rovníc, ktoré opisujú iba pohyby, nie túžby. Je to rozdiel medzi tvrdením, že stromu vyrastajú listy, pretože si to želá duch stromu, verzus skúmanie biochémie rastlín. Kognitívna veda preniesla boj proti nadprirodzenu do oblasti mysle.

Prečo je toto vynikajúca definícia nadprirodzena? Pre celý argument vás odkážem na Richarda Carriera. Ale uvážte toto: Predstavte si, že objavíte niečo, čo vyzerá ako *duch* prebývajúci v strome; dryáda, ktorá sa dokáže zhmotniť mimo stromu alebo v strome, a ktorá hovorí po anglicky o tom, že potrebuje chrániť svoj strom, a tak ďalej. A potom si predstavte, že zameriane mikroskop na túto stromovú vílu a ukáže sa, že sa skladá z častí – nie nejakých duchovných a nevýslovných častí, ako je tkanivo túžby a odev viery; ale skôr z takých častí ako kvarky a elektróny, častí, ktorých správanie je definované pomocou pohybu, nie myšlienok. Nebola by taká dryáda ihneď degradovaná do nudného katalógu všedných vecí?

Ale ak prijmeme definíciu nadprirodzena od Richarda Carriera, potom vzniká dilema: *chceme* dať náboženským tvrdeniam férovú príležitosť, ale zdá sa, že máme *veľmi dobré* dôvody na vylúčenie nadprirodzených vysvetlení *a priori*.

Chcem tým povedať, ako *by* vesmír vyzeral, keby redukcionizmus neplatil?

Už som definoval tézu redukcionizmu ako toto: ľudské mysle tvoria viacúrovňové *modely* skutočnosti, v ktorých sa vzory na vyššej úrovni a vzory na nižšej úrovni *reprezentujú* samostatne a explicitne. Fyzik pozná Newtonove rovnice gravitácie, Einsteinove rovnice gravitácie, a odvodenie toho prvého ako aproximácie toho druhého pri nízkych rýchlostiach. Ale tieto tri samostatné myšlienkové reprezentácie sú iba pomôckou pre ľudské poznávanie. Nie je to tak, že by *samotná skutočnosť* mala Einsteinove rovnice, ktorými sa riadia vysoké rýchlosti, Newtonove rovnice, ktorými sa riadia nízke rýchlosti, a „premosťovací zákon", ktorý vyhladzuje ich prechody. Skutočnosť samotná má iba jednu úroveň, Einsteinovskú gravitáciu. Je to iba Klam projekcie mysle, ktorý spôsobuje, že niektorí ľudia hovoria, akoby tie vyššie úrovne mohli mať samostatnú existenciu – rôzne úrovne organizácie môžu mať samostatné reprezentácie v ľudských mapách, ale samotné územie je jeden nerozdelený matematický objekt na nízkej úrovni.

Predstavme si, že by toto neplatilo.

Predstavme si, že by Klam projekcie mysle nebol klam, ale jednoducho pravda.

Predstavme si, že by lietadlo 747 malo základnú fyzikálnu existenciu oddelenú od kvarkov, z ktorých sa lietadlo 747 skladá.

Aké experimentálne pozorovania by ste očakávali, keby ste sa našli v takomto vesmíre?

Ak neviete prísť s dobrou odpoveďou na toto, nie je to *pozorovanie*, čo vylučuje „neredukcionistické" názory, ale a priori logická nekoherencia. Ak neviete povedať, aké predpovede robí „neredukcionistický" model, ako môžete povedať, že ho experimentálna indícia vyvracia?

Moja téza je, že neredukcionizmus je *zmätenie*; a keď si raz uvedomíte, že nejaký myšlienka je zmätenie, stane sa trochu zložitým predstaviť si, ako by vesmír vyzeral, keby toto zmätenie bola *pravda*.

Možno by som dostal nejaký viacúrovňový model sveta, a tento viacúrovňový model by zodpovedal jedna k jednej základným prvkom fyziky? Ale keď sú raz určené všetky tieto pravidlá, prečo by sa tento model jednoducho nesploštil do ďalšieho zoznamu základných vecí a ich interakcií? Musí sa všetko, čo dokážem v modeli *vidieť*, napríklad lietadlo 747 alebo ľudská myseľ, stať samostatnou skutočnou vecou? Ale čo ak vidím nejaký vzor v tomto novom supersystéme?

Supernaturalizmus je špeciálny prípad neredukcionizmu, kde neplatí, že lietadlo 747 je neredukovateľné, iba (niektoré) myšlienkové veci. Náboženstvo je špeciálny prípad supernaturalizmu, kde tie neredukovateľné myšlienkové veci sú Boh(ovia) a duše; a prípadne aj hriechy, anjeli, karma, atď.

Ak predpokladám existenciu mocnej bytosti so schopnosťou skúmať a meniť ľubovoľný prvok nášho pozorovaného vesmíru, avšak bytosti rozložiteľnej na nemyšlienkové časti, ktoré interagujú s prvkami nášho vesmíru zákonitým spôsobom; ak predpokladám, že táto bytosť chce nejaké konkrétne veci, ale „chce“ pomocou mozgu, ktorý sa skladá z častíc a polí; potom toto ešte nie je náboženstvo, iba naturalistická hypotéza o naturalistickom Matrixe. Keby sa zajtra oblaky rozostúpili a obrovská žiariaca beztvará postava by hromovým hlasom predniesla horeuvedený opis skutočnosti, nevyplývalo by z toho, že táto postava je nevyhnutne pravdovravná; ale mohol by som ukázať tieto zábery na hodinách fyziky a pokúsil by som sa z tejto teórie odvoliť testovateľné predpovede.

Naopak, náboženstvá ignorovali objav tej pradávnej netelesnej veci: všadeprítomnej v diele Prírody a skrytej v každom padajúcom liste: veľkej ako povrch planéty a starej miliardy rokov: samotnej nevytvorenej a zrodenej zo štruktúry fyziky: navrhujúcej bez mozgu a utvárajúcej všetok život na Zemi a mysle ľudstva. Prirodzený výber, keď ho Darwin predložil, nebol oslavovaný ako dlho očakávaný Stvoriteľ. Nebol *fundamentálne* myšlienkový.

Ale teraz sa dostávamek tej dileme: ak tradičné konvenčné normálne nudné chápanie fyziky a mozgu *je* správne, potom *z princípu* neexistuje spôsob, ako by si ľudská bytosť mohla konkrétne predstaviť alternatívny vesmír, v ktorom všetky veci *sú* neredukovateľne myšlienkové, a odvodiť o ňom testovateľné experimentálne predpovede. Pretože ak je ten nudný starý normálny model správny, váš mozog sa skladá z kvarkov, a tak si váš mozog bude vedieť predstaviť a správne predpovedať iba veci, ktoré možno predpovedať pomocou kvarkov. Budete vedieť zostaviť iba modely skladajúce sa z interagujúcich jednoduchých vecí.

Ľudia, ktorí žijú v redukcionistických vesmíroch si nedokážu správne predstaviť neredukcionistické vesmíry. Dokážu vysloviť slabiky „neredukcionistický“, ale nedokážu si to *predstaviť*.

Základná chyba antropomorfizmu, a dôvod, prečo nadprirodzené vysvetlenia znejú omnoho jednoduchšie než naozaj sú, je že váš mozog používam sám seba ako nepriehľadnú čiernu skrinku pri predpovedaní iných vecí označených ako „mysliace“. Keďže už máte veľké komplikované siete neurónových okruhov, ktoré implementujú vaše „chcenie“ vecí, pripadá vám, že môžete ľahko opísať, že voda „chce“ tiecť dole kopcom – jediné slovo „chce“ funguje ako páka, ktorá uvedie váš *vlastný* komplikovaný mechanizmus chcenia do obehu.

Alebo si predstavíte, že Boh má rád pekné veci a preto stvoril kvety. Váš vlastný okruh „krásy“ určuje, čo je „pekné“ a čo „nie pekné“. Nepoznáte však diagram svojich vlastných synapsií. Nedokážete opísať *nemyšlienkový* systém, ktorý vypočíta to isté označenie pre niečo, čo je „pekné“ alebo „nie pekné“ - nedokážete napísať počítačový program, ktorý by predpovedal vaše vlastné označenia. Ale toto je iba nedostatok poznania na vašej strane; neznamená to, že mozog nemá žiadne vysvetlenie.

Ak je „nudný pohľad“ na skutočnosť správny, potom *nikdy* nedokážete predpovedať nič neredukovateľné, pretože *vy* ste redukovateľní. Nikdy nemôžete dostať bayesovské potvrdenie hypotézy neredukovateľnosti, pretože *ľubovoľná predpoveď, ktorú dokážete urobiť*, je vďaka tomu niečím, čo by rovnako mohla predpovedať redukovateľná vec, konkrétne váš mozog.

Existujú koľaje, mimo ktorých naozaj *nedokážete* myslieť. Ak náš vesmír *naozaj je* turingovsky vypočítateľný, nikdy si nedokážeme *konkrétne* predstaviť niečo, čo nie je turingovsky vypočítateľné - bez ohľadu na to, *o* koľkých úrovniach veštcov zastavenia môžu hovoriť naši matematici, nedokážeme predpovedať, čo by nejaký veštec zastavenia naozaj *povedal*, takým spôsobom, aby sme to vedeli experimentálne odlíšiť od púheho vypočítateľného uvažovania.

Samozrejme, toto všetko predpokladá, že ten „nudný pohľad“ je správny. *Do tej miery*, do akej veríte, že evolúcia je pravdivá, nemali by ste očakávať, že sa stretnete so silnou indíciou proti evolúcii. Do tej miery, do akej veríte, že redukcionizmus je pravdivý, mali by ste očakávať, že neredukcionistické hypotézy budú *nesúvislé* a nesprávne. Do tej miery, do akej veríte, že supernaturalizmus je nepravdivý, mali by ste očakávať, že bude aj *nepredstaviteľný*.

Na druhej strane, ak sa nadprirodzená hypotéza ukáže byť pravdivá, potom pravdepodobne zároveň zistíte, že nie je nepredstaviteľná.

Takže sa oblúkom vráťme k otázke Inteligentného Dizajnu:

Mal by byť ID a priori vylúčený z experimentálnej falzifikácie a učební prírodovedy, pretože sa odvolávaním na nadprirodzené sám postavil mimo prírodnej filozofie?

Odpovedám: „Samozrejme, že nie.“ *Neredukovateľnosť* inteligentného dizajnéra nie je neodmysliteľná časť hypotézy ID. Pre každého neredukovateľného Boha, ktorého môžu zástancovia ID navrhovať, existuje zodpovedajúci redukovateľný mimozemšťan, ktorý sa správa v súlade s rovnakými predpoveďami – keďže samotní zástancovia ID sú redukovateľní; do tej miery, do akej verím, že redukcionizmus je naozaj správny, čomu verím do dosť veľkej miery, musím očakávať, že objavím redukovateľné formulácie všetkých údajne nadprirodzených modelov predpovedí.

Keby sme skúmali archeologické záznamy, aby sme otestovali tvrdenie, že Jehova rozdelil Červené More kvôli svojej explicitnej túžbe predviesť svoju nadľudskú silu, potom je malý rozdiel v tom, či je Jehova ontologicky základný, alebo je to mimozemšťan s nanotechnológiou, alebo Temný Pán Matrixu. Urobíte nejakú archeológiu, nenájdete naspodku Červeného Mora žiadne zvyšky kostier alebo brnenia, a naopak nájdete záznamy, že v tom čase Egypt ovládal väčšinu Kanaánu. Takže dáte na historický údaj v Biblii pečiatku „vyvrátené“ a idete ďalej. Daná hypotéza je súvislá, falzifikovateľná, a nesprávna.

Podobne je to s indíciami z biológie, že líšky sú navrhnuté, aby naháňali zajace, zajace sú navrhnuté, aby unikali líškam, a žiaden z nich nie je navrhnutý „pre dobro svojho druhu“ alebo „aby chránil harmóniu Prírody“; podobne s tým, ako sa je sietnica navrhnutá naopak, so svetlocitlivou stranou naspodku; a tak ďalej cez tisíc ďalších položiek indície pre rozbitý, nemorálny, nekompetentný dizajn. Jehovov model nášho mimozemského boha je súvislý, falzifikovateľný, a nesprávny – teda súvislý, pokiaľ vás nezaujíma, či je Jehova ontologicky základný, alebo je to iba mimozemšťan.

Skrátka premeňte nadprirodzenú hypotézu na jej zodpovedajúcu prirodzenú hypotézu. Skrátka urobte tie isté predpovede, tým istým spôsobom, bez tvrdenia, že nejaká myšlienková vec musí byť ontologicky základná. Ak treba, konzultujte s čiernou skrinkou svojho mozgu pri robení predpovedí – napríklad ak chcete hovoriť o „rozhnevanom bohu“ bez stavania plnohodnotnej rozhnevanej UI, ktorá by označovala správanie ako rozhnevané alebo nerozhnevané. Takto odvodíte predpovede, alebo sa pozriete na predpovede, ktoré robili starovekí teológovia bez znalosti našich experimentálnych výsledkov. Ak je experiment v rozpore s týmito predpoveďami, potom je férové povedať, že toto náboženské tvrdenie bolo vedecky vyvrátené. Dostalo svoju spravodlivú príležitosť na potvrdenie; je vylúčené a posteriori, nie a priori.

V konečnom dôsledku je redukcionizmus skrátka neviera vo *fundamentálne zložité* veci. Ak vám „fundamentálne zložité“ znie ako oxymoron... nuž, práve preto si myslím, že doktrína neredukcionizmu je *zmätenie*, a nie spôsob, akým veci mohli byť, ale nie sú. Lepšie je dávať si pozor, ak zistíte, že predpokladáte takéto veci.

Ale konečným pravidlom vedy je pozrieť a vidieť. Keby sa niekedy nejaký Boh objavil v hrome na hore, bolo by to niečo, na čo sa ľudia pozreli a videli.

Dodatok: Ľubovoľný údajný dizajnér Všeobecnej Umelej Inteligencie, ktorý hovorí o náboženských názoroch úctivým hlasom, očividne nie je odborníkom v redukovaní myšlienkových vecí na nemyšlienkové veci; a vôbec vie tak veľmi málo o úplných základoch, že je sotva uveriteľné, že by mohol byť odborníkom na toto umenie; pokiaľ to nie je dokonalý *idiot savant*. Alebo, samozrejme, pokiaľ vyslovene neklame. Tu nehovoríme o drobnej chybe.

*→

228. Nadprirodzené schopnosti

V predchádzajúcej kapitole som napísal:

> Ak je ten „nudný pohľad“ na skutočnosť správny, potom *nikdy* nedokážete predpovedať nič neredukovateľné, pretože *vy* ste redukovateľní. Nikdy nedokážete získať bayesovské potvrdenie pre hypotézu o neredukovateľnosti, pretože ľubovoľná *predpoveď, ktorú dokážete urobiť*, je vďaka tomu niečím, čo by mohla predpovedať aj nejaká redukovateľná vec, konkrétne váš mozog.

Benja Fallenstein komentoval:

> Myslím si, že aj keď v tomto prípade nikdy nedokážeš zostaviť empirický test, ktorého výsledok by mohol *logicky dokázať* neredukovateľnosť, neexistuje jasný dôvod veriť, že nedokážeš zostaviť test, ktorého kontrafaktuálny výsledok v neredukovateľnom svete by urobil neredukovateľnosť subjektívne *omnoho pravdepodobnejšou* (pri danom Occamovskom východisku).
>
> Bez zachádzania do redukovateľnosti/neredukovateľnosti, predstav si scenár, kde fyzikálny vesmír umožňuje postaviť hyperpočítač – ktorý robí napríklad operácie na ľubovoľných reálnych číslach – ale kde naše mozgy v skutočnosti toto nevyužívajú: možno ich dokonale simulovať pomocou obyčajných Turingových strojov...

Nuž, toto je veľmi inteligentný argument, Benja Fallenstein. Mám však na tvoj argument zdrvujúcu odpoveď, takú, že keď ju budeš počuť, okamžite sa vzdáš ďalšej debaty so mnou na túto konkrétnu tému:

Máš pravdu.

Žiaľ, nedostanem za toto bod za skromnosť, pretože po uverejnení včerajšieho článku som si sám uvedomil podobnú chybu – tentokrát týkajúcu sa Occamovej britvy a nadprirodzených schopností:

Ak sú názory a túžby neredukovateľné a ontologicky základné veci, alebo ak majú ontologicky základnú *zložku*, ktorú súčasná veda nezahŕňa, potom by bolo omnoho pravdepodobnejšie, že existuje nejaké ontologické pravidlo, ktorým sa riadi interakcia rôznych myslí – interakcia, ktoré presahuje bežné „materiálne“ prostriedky komunikácie, ako sú zvukové vlny, známe súčasnej vede.

Ak je naturalizmus pravdivý, potom existuje zodpovedajúci redukcionistický model, ktorý robí *rovnaké predpovede*, ako ľubovoľná konkrétna predpoveď, ktorú môže ľubovoľný parapsychológ povedať o telepatii.

Vlastne, ak je naturalizmus správny, potom jediný dôvod, prečo si vieme *predstaviť* názory ako „základné“ je vďaka nedostatku sebapoznania našich vlastných neurónov – vďaka osobitnej reflektívnej architektúre našej vlastnej mysle, ktorá nám odhaľuje triedu „názor“, ale skrýva mechanizmus za ním.

Napriek tomu, objav prenosu informácií medzi mozgami, v neprítomnosti žiadneho známeho hmotného spojenia medzi nimi, je *pravdepodobnostne* privilegovaná predpoveď nadprirodzených modelov (takých, ktoré obsahujú ontologicky základné myšlienkové veci). Pretože je v takom prípade omnoho *jednoduchšie* mať nový zákon týkajúci sa názorov v rôznych mysliach, v porovnaní s „nudným“ modelom, kde sú názory zložitými konštruktmi neurónov.

Nádej v nadprirodzené schopnosti vzniká z považovania názorov a túžob za dostatočne základné objekty, že až môžu mať *bezprostredné* spojenie so skutočnosťou. Ak sú názory vzory neurónov vyrobených zo známej hmoty, ktorých vstupy dodávajú orgány ako je oko, zostavené zo známej hmoty, a ktorých výstupy idú cez svaly vyrobené zo známej hmoty, a toto sa zdá dostatočné na vysvetlenie všetkých známych ľudských myšlienkových schopností, potom nie je dôvod očakávať niečo viac – žiaden dôvod predpokladať ďalšie spojenia. Preto redukcionisti neočakávajú nadprirodzené schopnosti. Preto by

→ http://lesswrong.com/lw/tv/excluding_the_supernatural/

pozorovanie nadprirodzených schopností bolo silnou indíciou pre nadprirodzeno v zmysle Richarda Carriera.

Máme Occamovo pravidlo, ktoré v modeli počíta počet ontologicky základných tried a ontologicky základných pravidiel, a penalizuje počet vecí. Ak je naturalizmus správny, potom snaha počítať „názor“ alebo „vzťah medzi názorom a skutočnosťou“ ako jednu základnú vec je jednoducho pomýlený antropomorfizmus; pokúša nás k tomu iba náhoda vnútornej architektúry nášho mozgu. Ale ak skrátka *pôjdete v súlade* s týmto pomýleným pohľadom, potom priradíte omnoho väčšiu pravdepodobnosť nadprirodzeným schopnostiam než im priraďuje naturalizmus, pretože dokážete nadprirodzené schopnosti implementovať pomocou napohľad jednoduchších zákonov.

Preto by skutočný objav nadprirodzených schopností naznačoval, že ľudsky naivné Occamovo pravidlo bolo v skutočnosti lepšie kalibrované než sofistikované naturalistické Occamovo pravidlo. Tvrdilo by to, že redukcionisti sa po celý čas mýlili, keď sa snažili rozobrať mozog; že to, čo naša myseľ ukazuje ako napohľad jednoduchú páku, naozaj bola jednoduchá páka. Naivní dualisti by boli mali pravdu od samotného začiatku, a preto by sa ich dávne želanie mohlo stať pravdou.

Takže telepatia, a schopnosť ovplyvňovať udalosti iba tým, že si ich želáme, a jasnovidectvo, by v prípade objavenia všetky boli silnou bayesovskou indíciou v prospech hypotézy, že názory sú ontologicky základné. Nie logický dôkaz, ale silná bayesovská indícia.

Ak je redukcionizmus správny, potom ľubovoľný sci-fi príbeh obsahujúci nadprirodzené schopnosti, môže byť výstupom systému s jednoduchými prvkami (napríklad mozog autora daného príbehu); ale ak *naozaj* objavíme nadprirodzené schopnosti, bolo by omnoho pravdepodobnejšie, že sa dejú veci, ktorí *naozaj* nemožno opísať redukcionistickými modelmi.

Z čoho vyplýva: Existencia nadprirodzených schopností je privilegované pravdepodobnostné tvrdenie neredukcionistických svetonázorov – je to *ich vlastná* predpoveď; oni ju vymysleli a predložili, navzdory redukcionistickým očakávaniam. Takže podľa zákonov vedy, ak sa objavia nadprirodzené schopnosti, neredukcionizmus vyhráva.

Preto mám dôveru v odmietanie nadprirodzených schopností ako a priori nedôveryhodných, napriek všetkým údajným experimentálnym indíciám v ich prospech.

* →

→ http://lesswrong.com/lw/tw/psychic_powers/

S: Kvantová fyzika a mnoho svetov

229. Kvantové vysvetlenia

Existuje rozšírená predstava, že kvantová mechanika *má* byť nepochopiteľná. To nie je dobré nastavenie mysle ani pre učiteľa, ani pre žiaka.

A mám skúsenosť, že témy povestné svojou „nepochopiteľnosťou“ často nie sú naozaj také zložité ako matematika, najmä aj chcete iba veľmi základné - ale stále matematické - chápanie toho, čo sa tam vlastne deje.

Nie som fyzik, a je známe, že fyzici neznášajú, keď niekto, kto nie je profesionálny fyzik, hovorí o kvantovej mechanike. Ale mám nejaké skúsenosti s vysvetľovaním matematických vecí, ktoré sa údajne „ťažko dajú pochopiť“.

Napísal som „Intuitívne vysvetlenie bayesovského uvažovania“, lebo sa ľudia sťažovali, že Bayesova veta je „kontraintuitívna“ - naozaj bola *slávna* svojou kontraintuitívnosťou – a to mi nepripadalo správne. Tá rovnica nevyzerala dosť zložito na to, aby sa zaslúžila takú obávanú povesť, akú mala. Skúsil som ju teda vysvetliť *po svojom* a nepodarilo sa mi dosiahnuť svoj pôvodný cieľ žiakov základnej školy, ale dostávam časté ďakovné maily od pôvodne popletených ľudí od novinárov po vysokoškolských učiteľov.

Okrem toho, ako Bayesovec, neverím v javy, ktoré sú *zo svojej podstaty* mätúce. Zmätok existuje v našich modeloch sveta, nie vo svete samotnom. Ak je nejaká oblasť všeobecne známa ako *mätúca*, nie iba *zložitá*... nemali by ste to nechať tak. To neuspokojuje; nie je dobré byť na takom mieste. Možno dokážete ten problém napraviť, možno nedokážete; ale nemali by ste byť *spokojní* s tým, že necháte študentov zmätených.

Prvá vec, v ktorej sa môj úvod bude líšiť od tradičného, štandarného úvodu do QM bude, že vám *nepoviem*, že kvantová mechanika má byť mätúca.

Nepoviem vám, že je okej, ak nerozumiete kvantovej mechanike, pretože kvantovej mechanike nerozumie nikto, ako raz tvrdil Richard Feynman. Existovalo historické obdobie, keď to bola pravda, ale my už v tej dobe nežijeme.

Nepoviem vám: „Kvantovej mechanike nemôžete rozumieť, môžete si na ňu iba zvyknúť.“ (Ako údajne povedal von Neumann; v tých temných desaťročiach, keď naozaj nikto *nerozumel* kvantovej mechanike.)

Vysvetlenia sú na to, aby vám urobili *menej zmätenými*. Ak máte pocit, že niečomu nerozumiete, to naznačuje *problém* – buď u vás alebo u vášho učiteľa – ale v každom prípade problém; a mali by ste sa rozhodnúť tento problém *vyriešiť*.

Nepoviem vám, že kvantová mechanika je *čudná, bizarná, mätúca* alebo *cudzia*. QM je kontraintuitívna, ale to je problém vašej intuície, nie problém kvantovej mechaniky. Kvantová mechanika tu bola miliardy rokov pred tým, než sa Slnko vytvorilo z medzihviezdneho prachu. Kvantová mechanika tu bola skôr než vy, a ak s tým máte problém, *vy* ste ten, kto sa musí zmeniť. QM sa iste nezmení. Neexistujú *prekvapujúce fakty*, iba *modely*, ktoré sú *prekvapené* z faktov; a ak je model prekvapený z faktov, nie je to preň dobré vysvedčenie.

Vždy je najlepšie myslieť na skutočnosť ako na dokonale normálnu. Od samotného začiatku sa nikdy nestala jediná nezvyčajná vec.

Cieľom je stať sa celkom doma v kvantovom vesmíre. Ako domorodec. Pretože tam naozaj žijete.

V tejto postupnosti o kvantovej mechanike budem konzistentne hovoriť, akoby kvantová mechanika bola *dokonale normálna*; a kde sa ľudské intuície odchýlia od kvantovej mechaniky, budem sa tým *intuíciám* posmievať, že sú čudné a nezvyčajné. To môže vyzerať čudne, ale pointa je otočiť vašu myseľ naopak, na *prirodzený* kvantový uhol pohľadu.

Ďalšia vec: Tradičný úvod do kvantovej mechaniky verne nasleduje poradie, v akom bola kvantová mechanika objavená.

Tradičný úvod začína tým, že povie, že hmota sa občas správa ako malé biliardové gule hopkajúce okolo, a niekedy sa správa ako hrebene a údolia pohybujúce sa jazerom vody. Potom tradičný úvod dá nejaké príklady, ako sa hmota správa ako malá biliardová guľa, a nejaké príklady, ako sa správa ako morská vlna.

Nuž, je historický fakt, že vtedy, keď študenti hmoty zisťovali všetky tieto veci a nemali *poňatie* o skutočnej matematike za tým, títo raní vedci si najprv mysleli, že hmota je ako malé biliardové gule. A potom, že je ako morské vlny. A potom opäť, že je ako biliardové gule. A potom boli títo raní vedci *naozaj* zmätení a zostali tak niekoľko desaťročí, dokiaľ sa to konečne nevyriešilo v druhej polovici dvadsiateho storočia.

Zaťahovať moderného študenta do tohto všetkého môže byť *historicky realistický* prístup k téme, ale zároveň to zabezpečuje historicky realistický výsledok *úplného zmätku*. Hovoriť ašpirujúcim mladým fyzikom o „časticovo vlnovej dualite“ je ako začať u študentov chémie so štyrmi živlami.

Elektrón *nie je* biliardová guľa, a *nie je* to hrebeň a údolie pohybujúce sa jazerom vody. Elektrón je matematicky odlišný druh veci, *vždy a za každých okolností*, a musíme ho prijať taký, aký sám je.

Vesmír nie je nerozhodný medzi používaním častíc a vĺn, neschopný rozmyslieť si, čo chce. To iba ľudské *intuície* ohľadom QM sa prepínajú tam a späť. Intuície, ktoré máme pre biliardové gule, a intuície, ktoré máme pre hrebene a údolia v jazere vody, obe vyzerajú, že by sa *trochu* dali použiť na elektróny, v rôznych prípadoch a za rôznych okolností. Ale pravda je, že obe intuície sú skrátka *nevhodné*.

Ak budete skúšať niektoré dni myslieť na elektrón, akoby na biliardovú guľu, a v iné dni ako na morskú vlnu, naozaj sa *popletiete do nevidím*.

Napriek tomu to vaše oči sú váhavé a nestabilné, nie samotný svet.

Ďalej:

Poradie, v ktorom ľudstvo *objavilo* veci nie je nevyhnutne tým najlepším poradím, v ktorom by sa mali *učiť*. Ľudia najprv zistili, že naokolo pobiehajú iné zvieratá. Potom sme ich rozsekali a zistili sme, že sú plné orgánov. Potom sme tie orgány starostlivo preskúmali a zistili sme, že sa skladajú z tkanív. Potom sme sa na tieto tkanivá pozreli pod mikroskopom a objavili sme bunky, ktoré sa skladajú z bielkovín a nejakých ďalších chemicky zostavených vecí. Ktoré sa skladajú z molekúl, tie sa skladajú z atómov, a tie sa skladajú z protónov a neutrónov a elektrónov, *ktoré sú omnoho jednoduchšie než celé zvieratá, ale boli objavené o desaťtisíce rokov neskôr*.

Fyzika nezačína rozprávaním o biológii. Prečo by teda mala začať rozprávať o komplikovaných javoch na veľmi vysokej úrovni, ako sú povedzme pozorované výsledky pokusov?

Zvyčajný spôsob vyučovania QM stále kladie dôraz na experimentálne výsledky. Ja teda rozumiem, že to z racionalistického pohľadu znie pekne. Verte mi, že tomu rozumiem.

Ale zdá sa mi, že výsledom je dotiahnutie velkých zložitých matematických nástrojov, ktoré potrebujete na analýzu situácií zo skutočného sveta, skôr než študent pochopí, čo sa *v skutočnosti* odohráva v tých najjednoduchších prípadoch.

Je to ako keby sme učili programátorov, ako písať konkurentné mnohovláknové programy skôr než vedia, ako sčítať dve premenné, pretože konkurentné mnohovláknové programy sú bližšie ku každodennému životu. Byť bližšie ku každodennému životu nie je vždy silným odporúčaním na prvú lekciu.

Možno takéto monomaniakálne sústredenie na experimentálne pozorovania dávalo zmysel v temných desaťročiach, keď *nikto* nerozumel, čo sa tu vlastne deje, tak ste tam *nemohli* začať, a všetky vaše modely boli iba tajomná matematika, ktorá dávala dobré experimentálne predpovede... tento uhol pohľadu na kvantovú fyziku stále nájdete vykreslený v mnohých knihách... ale možno dnes je hodno skúsiť iný uhol? Výsledkom štandardného prístupu je štandardný zmätok.

Klasický svet je striktne zahrnutý v kvantovom svete, ale pozeranie z klasického pohľadu robí všetko väčším a zložitejším. Každodenný život je vyššia úroveň organizácie, ako molekuly verzus kvarky – obrovské katalógy molekúl, šesť kvarkov. Myslím si, že sa oplatí učiť najprv z pohľadu kvantového sveta, a o klasických experimentálnych výsledkoch hovoriť dodatočne.

Nezačnem s normálnym klasickým svetom, aby som potom hovoril o bizarnom kvantovom pozadí skrytom v zákulisí. Kvantový svet *je* javisko, a definuje normálnosť.

Nebudem hovoriť, akoby klasický svet bol skutočný život, a akoby klasický svet občas odoslal požiadavku na experimentálny výsledkok serveru kvantovej fyziky, a server kvantovej fyziky urobil nejaké nezvyčajné výpočty a odoslal naspäť klasický experimentálny výsledok. Budem hovoriť, akoby kvantový svet bol naozaj naozajstný a klasický svet niečo veľmi vzdialené. Nie iba pretože je tak ľahšie byť domorodcom v kvantovom vesmíre, ale pretože, v jadre veci, toto je pravda.

Nakoniec, budem zaujímať striktne realistický pohľad na kvantovú mechaniku – kvantový svet naozaj existuje, naše rovnice opisujú územie a nie naše mapy územia, a klasický svet existuje iba implicitne vnútri toho kvantového. Nebudem diskutovať o nerealistických pohľadoch v prvých etapách môjho úvodu, akurát poviem, že by ste sa nemali nechať popliesť niektorými intuíciami, ktoré nerealisti používajú ako podporu. Nebudem sa za toto ospravedlňovať, a rád by som poprosil všetkých nerealistov ohľadom kvantovej mechaniky, aby počkali a zdržali sa komentárov, kým ich nevyzvem v neskoršom článku. Urobte mi túto láskavosť, prosím. Myslím si, že neralizmus je jedna z hlavných vecí, ktoré pletú perspektívnych študentov, a bránia im v schopnosti konkrétne si predstaviť kvantové javy. O tejto téme *budem* diskutovať explicitne v neskoršom článku.

Ale každý by si mal uvedomovať, že aj keď nebudem o tejto veci diskutovať od začiatku, existuje početná komunita vedcov, ktorí spochybňujú realistický pohľad na QM. Ja osobne si nemyslim, že je hodno skúmať obe cesty; som čistý realista z dôvodov, ktoré budú zrejmé. Ale ak si čítate môj úvod, dostanete môj pohľad. Nie je to iba môj pohľad. Je to pravdepodobne väčšinový pohľad medzi teoretickými fyzikmi, ak to niečo znamená (hoci budem o záležitosti argumentovať oddelene od prieskumov verejnej mienky). Napriek tomu, nie je to jediný existujúci pohľad v komunite moderných fyzikov. Necítim sa povinný prezentovať zvyšné pohľady *hneď na úvod*, ale cítim sa povinný upozorniť mojich čitateľov, že iné pohľady *existujú*, a že ich nebudem prezentovať počas úvodných častí tohto úvodu.

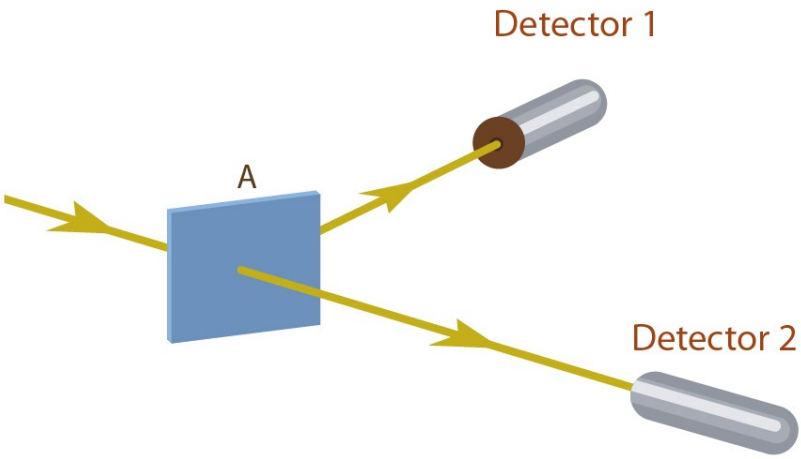
Aby som to zhrnul, mojím cieľom je naučiť vás myslieť ako *domorodec v kvantovom vesmíre*, nie ako *zdráhavý turista*.

Prijmite skutočnosť. Pevne si ju pritisnite.

*→

230. Konfigurácie a amplitúdy

Vesmír sa teda neskladá z malých biliardových gúľ, ani sa neskladá z hrebeňov a údolí v jazere éteru... Z akých vecí sa teda veci skladajú?



Obrázok 230.1

→ http://lesswrong.com/lw/pc/quantum_explanations/

Na obrázku 230.1 vidíme v bode A polopriepustné zrkadlo a dva detektory fotónov, 1 a 2.

Keď raní vedci robili takéto pokusy, začali byť zmätení ohľadom toho, čo tie výsledky znamenajú. Poslali fotón smerom k polopriepustnému zrkadlu, a v polovici prípadov videli cvaknúť detektor v 1, a v druhej polovici prípadov videli cvaknúť detektor v 2.

Títo raní vedci – tomuto sa budete smiať – si mysleli, že to strieborné zrkadlo v polovici prípadov odrážalo fotón, a v polovici prípadov ho nechávalo prejsť.

Ha, ha! Akoby polopriepustné zrkadlo za rôznych okolností robilo rôzne veci! Chcem, aby ste sa tejto predstavy vzdali, pretože ak sa upnete na to, čo si títo raní vedci mysleli, budete extrémne zmätení. To polopriepustné zrkadlo sa vždy správa podľa rovnakého pravidla.

Keby ste išli napísať počítačový program, ktorý bol tento pokus – nie počítačový program, ktorý *predvída* výsledky tohto pokusu, ale počítačový program, ktorý pripomína skutočnosť za tým – mohol by vyzerať asi takto:

Na začiatku programu (na začiatku pokusu, na začiatku času) existuje istá matematická vec, zvaná *konfigurácia*. Môžete na túto konfiguráciu myslieť ako na zodpovedajúcu stavu: „Jeden fotón mieri zo zdroja fotónov smerov k polopriepustnému zrkadlu“ alebo iba „Fotón smerom do A.“

Konfigurácia môže obsahovať jednu komplexnú hodnotu - „komplexnú“ v zmysle komplexných čísel (a + bi), kde i je definované ako odmocnina z -1. Na začiatku programu sa v konfigurácii „Fotón smerom do A“ už nachádza nejaké komplexné číslo. Presná hodnota nás nezaujíma, pokiaľ to nie je nula. Nech má teda konfigurácia „Fotón smerom do A“ hodnotu (-1 + 0i).

Toto všetko je fakt na území, nie popis niekoho vedomostí. Konfigurácia nie je návrh alebo možnosť alebo spôsob, ako by svet mohol vyzerať. Konfigurácia je premenná v programe – môžete na ňu myslieť ako na miesto v pamäti, ktorého index je „Fotón smerom do A“ - a nachádza sa tam vonku, v území.

Keďže komplexné čísla priradené ku konfiguráciám nie sú kladné celé čísla medzi 0 a 1, nehrozí, že by sme si ich poplietli s pravdepodobnosťami. „Fotón smerom do A“ má komplexnú hodnotu -1, čo je ťažké vidieť ako stupeň viery. Komplexné čísla sú hodnoty v programe, opäť, tam vonku v území. Tieto komplexné čísla budeme volať *amplitúdy*.

Existujú dve ďalšie konfigurácie, ktoré nazveme „Fotón smerom z A do detektoru 1“ a „Fotón smerom z A do detektoru 2.“ Tieto konfigurácie zatiaľ nemajú komplexnú hodnotu; priradí sa im počas behu programu.

Ideme počítať amplitúdy „Fotón smerom z A do 1“ a „Fotón smerom z A do 2“ pomocou hodnoty „Fotón smerom z A“ a pravidla, ktoré opisuje polopriepustné zrkadlo v A.

Približne povedané, pravidlo pre polopriepustné zrkadlo znie: „Vynásob 1, keď fotón ide priamo, a vynásob i, keď fotón odbočí v prvom uhle.“ Toto je všeobecné pravidlo, ktoré sa spája amplitúdu konfigurácie „fotón ide dnu“ s amplitúdami, ktoré idú do konfigurácií „fotón vychádza priamo“ a „fotón sa odráža“.[207]

Napojíme teda amplitúdu konfigurácie „Fotón smerom do A“, ktorá je (-1 + 0i), do polopriepustného zrkadla v A, a toto vyšle amplitúdu (-1 + 0i) × i = (0 + -i) do „Fotón smerom z A do 1“, a vyšle aj amplitúdu (-1 + 0i) × 1 = (-1 + 0i) do „Fotón smerom z A do 2.“

V pokuse na obrázku 230.1 sú toto všetky konfigurácie a všetky prenášané amplitúdy, o ktoré sa potrebujeme starať, takže sme skončili. Prípadne, ak chcete myslieť na „Detektor 1 má fotón“ a „Detektor 2 má fotón“ ako na samostatné konfigurácie, tak tie jednoducho zdedia svoje hodnoty z „A do 1“ a „A do 2“. (V skutočnosti by sa tieto zdedené hodnoty mali vynásobiť ďalším komplexným činiteľom zodpovedajúcim vzdialenosti od A do detektora, ale dnes budeme toto ignorovať, a budeme

207 [207] **Poznámka redaktora:** Prísne povedané, štandardné polopriepustné zrkadlo by dávalo pravidlo „vynásob -1, keď fotón odbočí v pravom uhle“, nie „vynásob i“. Základný scenár, ktorý tu autor opísal, nie je fyzikálne nemožný, a jeho použitie nemá vplyv na podstatu argumentu. Avšak študentov fyziky môže mýliť, keď porovnajú túto diskusiu s diskusiami v učebniciach o Machových-Zehnderových interferometroch. Ponechávame tento autorov svojský prístup v texte, pretože odstraňuje potrebu upresňovať, ktorá strana zrkadla je napoly postriebrená, čím sa experiment zjednodušuje.

predpokladať, že všetky vzdialenosti prejdené v našom pokuse zhodou okolností zodpovedajú komplexnému činiteľu 1.)

Výsledný stav programu je teda:

Konfigurácia „Fotón smerom do A“: (-1 + 0i)

Konfigurácia „Fotón smerom z A do 1“: (0 + -i)

Konfigurácia „Fotón smerom z A do 2“: (-1 + 0i)

a prípadne

Konfigurácia „Detektor 1 má fotón“: (0 + -i)

Konfigurácia „Detektor 2 má fotón“: (-1 + 0i)

Ten istý výsledok dostanete – rovnaké amplitúdy uložené v rovnakých konfiguráciách – vždy keď zbehnete program (vždy keď vykonáte pokus).

Z *komplikovaných* dôvodov, do ktorých sa dnes nebudeme púšťať – sú to veci, ktoré patria na vyššiu úroveň organizácie bež základná kvantová mechanika, rovnako ako sú atómy omnoho zložitejšie než kvarky – neexistuje žiaden *jednoduchý* merací prístroj, ktorý by nám mohol priamo povedať presné amplitúdy každej konfigurácie. Nemôžeme priamo vidieť stav programu.

Ako teda fyzici vedia, čo sú tie amplitúdy?

Máme čarovný merací nástroj, ktorý nám vie povedať *druhú mocninu absolútnej hodnoty* amplitúdy danej konfigurácie. Ak je pôvodná komplexná amplitúda (a + bi), dostaneme kladné reálne číslo (a^2 + b^2). Spomeňte si na Pytagorovu vetu: ak si predstavíte komplexné číslo ako malú šípku zo začiatku súradníc dvojrozmernej roviny, potom nám tento čarovný nástroj povie druhú mocninu dĺžky tejto malej šípky, ale nepovie nám, ktorým smerom táto šípka ukazuje.

Aby sme boli presnejší, tento čarovný prístroj nám vlastne povie iba *pomery* druhých mocnín dĺžok amplitúd v nejakých konfiguráciách. Nevieme, aké dlhé sú tieto šípky v absolútnom zmysle, iba aké dlhé sú v porovnaní jedna s druhou. Ale ukazuje sa, že to je dosť informácie na to, aby sme dokázali rekonštruovať zákony fyziky – pravidlá programu. A tak môžem hovoriť o amplitúdach, namiesto iba o pomeroch druhých mocnín absolútnych hodnôt.

Keď zamávame týmto čarovným nástrojom nad „Detektor 1 má fotón“ a „Detektor 2 má fotón“, zistíme, že tieto konfigurácie majú rovnakú druhú mocninu absolútnej hodnoty – dĺžky šípok sú rovnaké. Tak povie čarovný nástroj. Robením *zložitejších* pokusov (čoskoro uvidíte) môžeme povedať, že tie pôvodné komplexné čísla mali pomer i k 1.

A čo je tento čarovný merací nástroj?

Nuž, z pohľadu každodenného života – veľmi, veľmi, veľmi vysoko nad kvantovou úrovňou a omnoho zložitejšie – ten čarovný merací nástroj je, že pošleme niekoľko fotónov smerom k polopriepustnému zrkadlu, jeden za druhým, a spočítame, koľko fotónov príde do detektora 1 verzus do detektora 2 pri pár tisíc pokusoch. Pomer medzi týmito hodnotami je pomer druhých mocnín absolútnych hodnôt amplitúd. Ale o dôvodoch, prečo to tak je, zatiaľ *nebudeme* uvažovať. Najprv sa naučíme chodiť, až potom behať. Nie je možné pochopiť, čo sa deje *celkom hore* na úrovni každodenného života, kým nepochopíte, čo sa deje v omnoho ľahších prípadoch.

Na dnes nám stačí, že máme čarovný merač pomeru druhých mocnín absolútnych hodnôt. A tento čarovný nástroj nám hovorí, že malá dvojrozmerná šípka pre konfiguráciu „Detektor 1 má fotón“ má rovnakú druhú mocninu dĺžky ako „Detektor 2 má fotón“. To je všetko.

Možno sa čudujete: „Keď ten čarovný nástroj funguje takto, čo nás motivuje používať kvantovú teóriu namiesto myslenia si, že polopriepustné zrkadlo odráža fotón v polovici prípadov?“

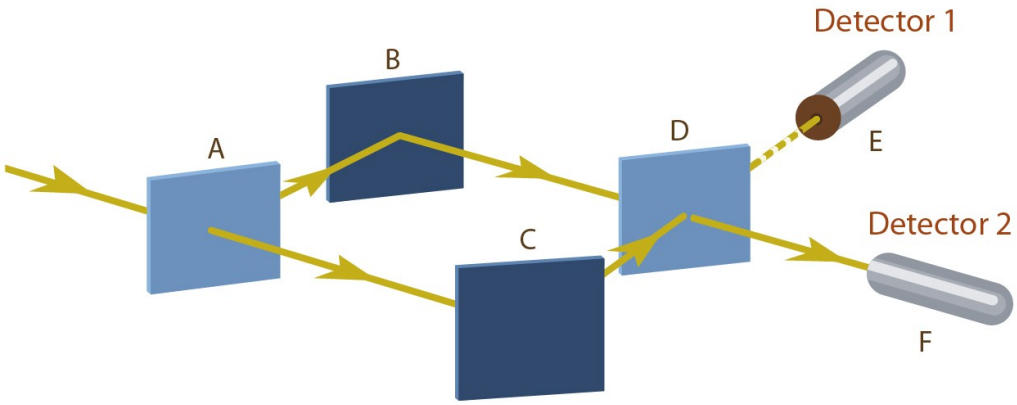
Nuž, to by ste si iba koledovali o zmätok – stavaním sa do historicky realistického stavu mysle a používaním každodenných intuícií. Hovoril som vari niečo o malej biliardovej guli, ktorá ide jedným alebo druhým smerom a možno sa odráža od zrkadla? Takto skutočnosť nefunguje. *Skutočnosť* sa týka komplexných amplitúd prúdiacich medzi konfiguráciami, a zákony tohto prúdenia sú pevne dané.

Ale ak trváte na tom, že chcete vidieť zložitejšiu situáciu, ktorú biliardový spôsob myslenia nedokáže spracovať, tu je zložitejší pokus:

Na obrázku 230.2 sú B a C úplné zrkadlá, a A a D sú polovičné zrkadlá. Čiara z D do E je prerušovaná z dôvodov, ktoré budú čoskoro zrejmé, ale amplitúda z D do E prúdi podľa celkom rovnakých zákonov.

Použime teda pravidlá, ktoré sme sa už naučili:

Na začiatku času má „Fotón smerom do A“ amplitúdu (-1 + 0i).



Obrázok 230.2

Pokračujeme spočítaním amplitúdy konfigurácií „Fotón smerom z A do B“ a „Fotón smerom z A do C“.

„Fotón smerom z A do B“ = i × „Fotón smerom do A“ = (0 + -i)

Podobne,

„Fotón smerom z A do C“ = 1 × „Fotón smerom do A“ = (-1 + 0i)

Úplné zrkadlá sa správajú (ako by človek čakal) ako polovica polopriepustného zrkadla – úplné zrkadlo iba ohne veci o pravý uhol a vynásobí ich i. (Aby som to vyjadril trochu presnejšie: Pri úplnom zrkadle sa vstupná amplitúda konfigurácie prichádzajúceho fotónu vynásobí činiteľom i a dostaneme konfiguráciu fotónu vychádzajúceho pod pravým uhlom.)

Takže:

„Fotón smerom z B do D“ = i × „Fotón smerom z A do B“ = (1 + 0i)

„Fotón smerom z C do D“ = i × „Fotón smerom z A do C“ = (0 + -i)

Kde „z B do D“ a „z C do D“ sú dve rôzne konfigurácie – nepíšeme jednoducho „fotón v D“ - pretože fotóny v týchto dvoch rôznych konfiguráciách prichádzajú v dvoch rôznych uhloch. A čo urobí D s fotónom, závisí od toho, v ktorom uhle fotón prichádza.

Opäť, pravidlo (voľne povedané) znie tak, že keď polopriepustné zrkadlo ohne svetlo v pravom uhle, amplitúda prúdiaca z konfigurácie prichádzajúceho fotónu do konfigurácie odchádzajúceho fotónu, je amplitúda konfigurácie prichádzajúceho fotónu vynásobená i. A keď sú dve konfigurácie vo vzťahu, že polopriepustné zrkadlo nechá svetlo prejsť priamo, amplitúda prúdiaca z konfigurácie prichádzajúceho fotónu sa vynásobí 1.

Takže:

- Z konfigurácie „Fotón smerom z B do D“ s pôvodnou amplitúdou (1 + 0i)
  - Amplitúda (1 + 0i) × i = (0 + i) prúdi do „Fotón smerom z D do E“
  - Amplitúda (1 + 0i) × 1 = (1 + 0i) prúdi do „Fotón smerom z D do F“.
- Z konfigurácie „Fotón smerom z C do D“ s pôvodnou amplitúdou (0 + -i)
  - Amplitúda (0 + -i) × i = (1 + 0i) prúdi do „Fotón smerom z D do F“
  - Amplitúda (0 + -i) × 1 = (0 + -i) prúdi do „Fotón smerom z D do E“.

Preto:

- *Celková* amplitúda prúdiaca do konfigurácie „Fotón smerom z D do E“ je (0 + i) + (0 + -i) = (0 + 0i) = 0.
- Celková amplitúda prúdiaca do konfigurácie „Fotón smerom z D do F“ je (1 + 0i) + (1 + 0i) = (2 + 0i).

(Môžete si to skúsiť prepočítať sami perom na papieri, ak ste sa niekde stratili.)

Záver je, že z „experimentálneho“ pohľadu na super vysokej úrovni, ktorú vnímame ako normálny život, neuvidíme *žiadne* fotóny namerané v E. Každý fotón vyzerá, že skončí v F. Pomer medzi druhými mocninami absolútnych hodnôt „z D do E“ a „z D do F“ je 0 k 4. Preto je čiara z D do E na tomto obrázku nakreslená prerušovane.

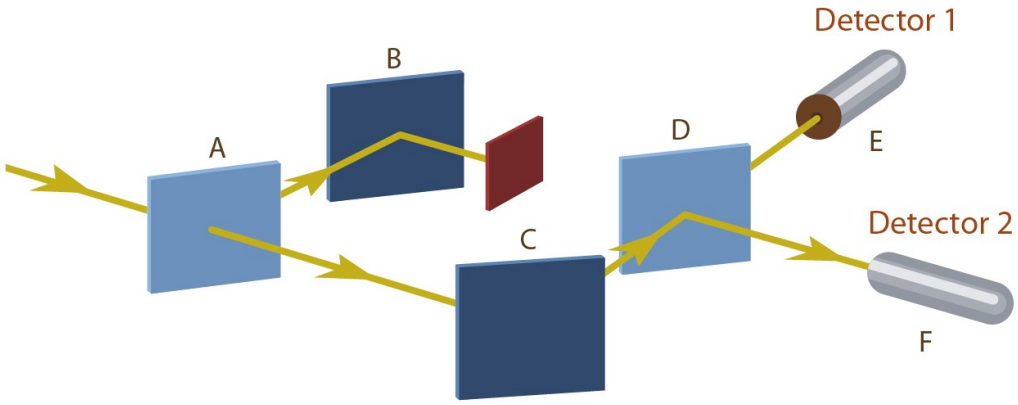
Toto nie je niečo, čo by sa dalo vysvetliť, keď si predstavujete, že popopriepustné zrkadlá odrážajú malé biliardové gule v polovici prípadov. *Musíte* na to myslieť pomocou tokov amplitúdy.

Keby polopriepustné zrkadlá odrážali malú biliardovú guľu v polovici prípadov, pri tomto nastavení by malá guľa skončila asi v polovici prípadov v detektore 1, a asi v polovici prípadov v detektore 2. Čo sa nedeje. Preto takto nerozmýšľajte.

Možno poviete: „Počkaj chvíľu! Viem si predstaviť inú hypotézu, ktorý vysvetľuje tento výsledok. Čo ak, keď polopriepustné zrkadlo odráža fotón, urobí s ním niečo, čo spôsobí, že sa druhýkrát neodrazí? A keď prepustí fotón priamo, urobí s ním niečo, aby sa nabudúce odrazil.“

Nuž, naozaj nie je dôvod, aby boli pravidlá takéto komplikované. Pamätajte na Occamovu britvu. Zostaňme pri jednoduchých, normálnych tokoch amplitúdy medzi konfiguráciami.

Ale ak chcete *ďalší* pokus, ktorý vyvráti vašu *novú* alternatívnu hypotézu, je na obrázku 230.3:



Obrázok 230.3

Tu sme nechali celé nastavenie pokusu rovnaké, iba sme dali malý zavadzajúci predmet medzi B a D. To zabezpečuje, že amplitúda „Fotón smerom z B do D“ je 0.

Keď odstránite príspevky amplitúdy z tejto konfigurácie, skončíte s výsledkami (1 + 0i) v „Fotón smerom z D do F“ a (0 + -i) v „Fotón smerom z D do E“.

Druhé mocniny absolútnych hodnôt (1 + 0i) a (0 + -i) sú obe 1, takže čarovný merací nástroj by nám mal povedať, že pomer druhých mocním absolútnych hodnôt je 1. Späť hore na úrovni, kde existujú fyzici, mali by sme zistiť, že v polovici prípadov cvakne detektor 1, a v polovici prípadov detektor 2.

To isté sa stane, ak dáme prekážku medzi C a D. Amplitúdy sú iné, ale pomer druhých mocnín absolútnych hodnôt je stále 1, takže v polovici prípadov cvakne detektor 1, a v polovici prípadov cvakne detektor 2.

Toto by sa *nemohlo* stať s malou biliardovou guľou, ktorá sa v polopriepustnom zrkadle buď odrazí alebo neodrazí.

Pretože komplexné čísla môžu mať opačný smer, ako 1 a -1, alebo i a -i, toky amplitúdy sa môžu navzájom zrušiť. Amplitúda prúdiaca z konfigurácie X do konfigurácie Y môže byť zrušená rovnako veľkou a opačne orientovanou amplitúdou prúdiacou z konfigurácie Z do konfigurácie Y. V skutočnosti sa presne toto deje v tomto pokuse.

V teórii pravdepodobnosti, keď sa niečo môže stať buď jedným alebo druhým spôsobom, X alebo ~X, potom P(Z) = P(Z | X) × P(X) + P(Z | ~X) × P(~X). A všetky pravdepodobnosti sú kladné. Ak teda zistíte, že pravdepodobnosť, že sa stane Z za predpokladu X je 1/2, a že pravdepodobnosť, že sa stane X je 1/3, potom celková pravdepodobnosť, že sa stane Z je *prinajmenšom* 1/6 bez ohľadu na to, *čo* sa deje v prípade ~X. Neexistuje nič také ako záporná pravdepodobnosť, menej než nemožné, alebo dôveryhodnosť (0 + i), preto sa *stupne presvedčenia* nemôžu navzájom vykrátiť tak, ako amplitúdy.

Navyše, pravdepodobnosť existuje iba v mysli; a my teraz hovoríme *o* území, o programe, ktorý je skutočnosť, nie *o* ľudskom poznávaní ani stavoch čiastočného poznania.

Rovnako, konfigurácie nie sú *predpoklady*, ani *výroky*, ani *spôsoby, ako by svet mohol byť*. Konfigurácie nie sú jazykové konštrukty. Prídavné mená ako *pravdepodobný* a *možný* sa na ne nevzťahujú; nie sú to názory ani vety ani možné svety. Nie sú *pravdivé* alebo *nepravdivé*, ale jednoducho *skutočné*.

V pokuse na obrázku 230.2 sa nenechajte zviesť k mysleniu niečoho ako: „Fotón ide buď do B alebo do C, ale *mohol* by ísť ľubovoľnou cestou, a táto pravdepodobnosť interferuje s jeho schopnosťou ísť do E...“

Nedáva zmysel predstavovať si, ako niečo, čo „sa mohlo stať, ale nestalo sa“ pôsobí na svet. Môžeme si *predstavovať* veci, ktoré sa mohli stať, ale nestali – napríklad pomyslieť si „Ach, to auto ma skoro zrazilo“ - a naša predstavivosť môže pôsobiť na naše budúce správanie. Ale udalosť predstavovania si je skutočná udalosť, ktorá sa naozaj stala, a *to* je to, čo pôsobí. Je to vaša predstava neskutočnej udalosti – vaše celkom skutočná predstava, implementovaná v celkom fyzikálnom mozgu – čo ovplyvňuje vaše správanie.

Predstavovať si, že *samotná udalosť*, že vás auto zrazilo – táto udalosť, ktorá sa vám mohla stať, ale v skutočnosti sa nestala – priamo *kauzálne* pôsobí na vaše správanie, to je mýlenie si mapy s územím.

To, čo ovplyvňuje svet, je skutočné. (Keby veci mohli ovplyvňovať svet bez toho, že by boli „skutočné“, nebolo by jasné, čo vlastne slovo „skutočné“ znamená.) Konfigurácie a toky amplitúd sú príčiny, a majú viditeľné následky; sú skutočné. Konfgurácie nie sú možné svety, ani amplitúdy nie sú stupne viery, o nič viac ako vaša stolička nie je možný svet a obloha nie je stupeň viery.

Čo teda konfigurácia *je*?

Tak o tomto získate jasnejšiu predstavu v nasledujúcich článkoch.

Ale aby som vám dal rýchlu predstavu, ako sa skutočný obrázok líši od zjednodušenej verzie, ktorú sme videli dnes...

Náš pokus sa týkal iba jednej pohybujúcej sa častice, jediného fotónu. Skutočné konfigurácie sa týkajú mnohých častíc. Zajtrajší článok sa bude vedomovať prípadu viac ako jednej častice, a to by vám malo dať omnoho jasnejšiu predstavu o tom, čo je to konfigurácia.

Každá konfigurácia, o ktorej sme hovorili, by *mala* popisovať dohromady polohy všetkých častíc v zrkadlách a detektoroch, nie iba polohu jedného fotónu hopkajúceho okolo.

V skutočnosti, tie *naozaj ozajstné* konfigurácie sú dohromady pre polohy všetkých častíc vo vesmíre, vrátane častíc, z ktorých sa skladajú experimentátori. Možno vidíte, prečo si odkladám pojem *experimentálneho výsledku* na neskoršie články.

V skutočnom svete je amplitúda spojitou distribúciou v spojitom *priestore* konfigurácií. Dnešné „konfigurácie“ boli kockaté a digitálne, takisto aj naše „toky amplitúdy“. Bolo to, akoby sme hovorili o fotóne, ktorý sa teleportuje z jedného miesta na druhé.

Dostaneme sa k atómom a molekulám a ľuďom a všetkým takýmto veciam z diferencovateľných distribúcií amplitúd v spojitom konfiguračnom priestore, *neskôr*.

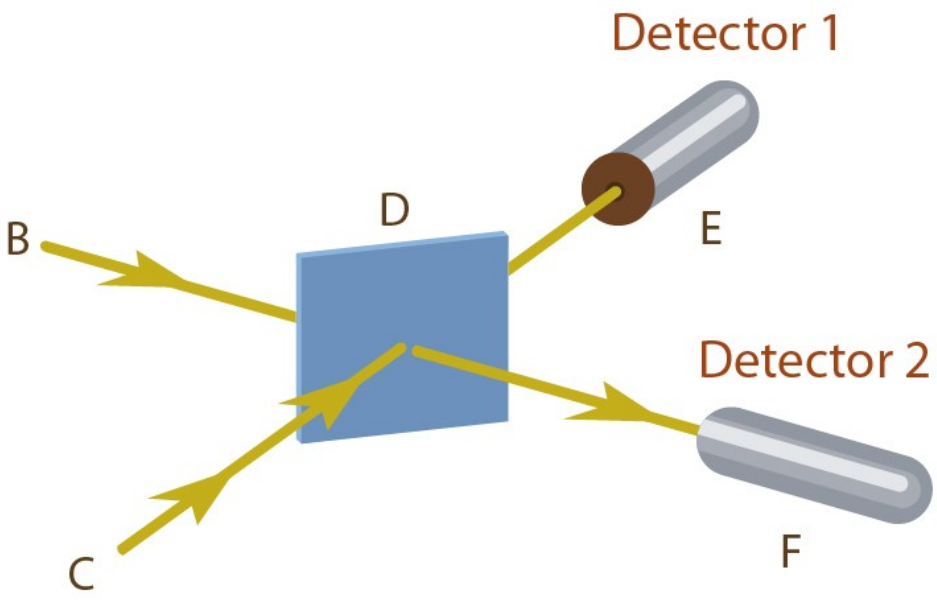
Ak vám nič z toho nedáva zmysel, nebojte sa. Vyjasní sa to v nasledujúcich článkoch. Len som vám chcel dať nejakú predstavu, kam toto smeruje.

*→

→ http://lesswrong.com/lw/pd/configurations_and_amplitude/

231. Spoločné konfigurácie

Kľúčom k pochopeniu konfigurácií a teda kľúčom k pochopeniu kvantovej mechaniky, je uvedomenie si na naozaj inštinktívnej úrovni, že konfigurácie sa týkajú viac než jednej častice.



Obrázok 231.1

Ako pokračovanie k predchádzajúcej kapitole, obrázok 231.1 ukazuje zmenenú verziu pokusu, kde sme smerom k D posielame *dva* fotóny naraz, zo zdrojov B a C.

Začiatočná konfigurácia je teda:

Fotón smerom z B do D,

a fotón smerom z C do D.

Opäť, povedzme, že táto štartová konfigurácia má amplitúdu (-1 + 0i).

A pamätajte, že pravidlo polopriepustného zrkadla (v D) znie, že odraz v pravom uhle násobí i, a priame prejdenie násobí 1.

Toky amplitúdy zo začiatočnej konfigurácie, samostatne zvažujúc štyri prípady odrazenia a neodrazenia každého fotónu, sú teda:

1. Fotón „z B do D“ sa odraží a fotón „z C do D“ sa odrazí. Táto amplitúda prúdi do konfigurácie „Fotón smerom z D do E, a fotón smerom z D do F“. Tento tok amplitúdy je (-1 + 0i) × i × i = (1 + 0i).
2. Fotón „z B do D“ sa odrazí a fotón „z C do D“ prejde rovno. Táto amplitúda prúdi do konfigurácie „Dva fotóny smerom z D do E.“ Tento tok amplitúdy je (-1 + 0i) × i × 1 = (0 + -i).
3. Fotón „z B do D“ prejde rovno a fotón „z C do D“ sa odrazí. Táto amplitúda prúdi do konfigurácie „Dva fotóny smerom z D do F.“ Tento tok amplitúdy je (-1 + 0i) × 1 × i = (0 + -i).
4. Fotón „z B do D“ prejde rovno a fotón „z C do D“ prejde rovno. Táto amplitúda prúdi do konfigurácie „Fotón smerom z D do F, a fotón smerom z D do E.“ Tento tok amplitúdy je (-1 + 0i) × 1 × 1 = (-1 + 0i).

Lenže – a toto je *veľmi dôležitá a základná myšlienka kvantovej mechaniky* – amplitúdy v prípadoch 1 a 4 prúdia do *tej istej* konfigurácie. Či fotón z B a fotón z C prejdú oba rovno, alebo sa oba odrazia, výsledná konfigurácia je *jeden fotón smerom k E a druhý fotón smerom k F*.

Sčítame teda tieto dva toky prichádzajúcej amplitúdy z prípadu 1 a prípad 4, a dostaneme celkovú amplitúdu (1 + 0i) + (-1 + 0i) = 0.

Keď zamávame naším čarovným čítačom druhej mocniny absolútnej hodnoty nad troma výslednými konfiguráciami, zistíme, že „Dva fotóny v detektore 1“ a „Dva fotóny v detektore 2“ majú

rovnaké druhé mocniny absolútnych hodnôt, ale „Fotón v detektore 1, a fotón v detektore 2“ má druhú mocninu absolútnej hodnoty nula.

Omnoho vyššie nad úrovňou experimentu, nikdy nezistíme, že by detektor 1 a detektor 2 cvakli oba naraz. Zistíme, že detektor 1 cvakol dvakrát, alebo že detektor 2 cvakol dvakrát, rovnako často. (Za predpokladu, že som sa tu v matematike ani vo fyzike nepomýli. Nerobil som tento pokus naozaj.)

Totožnosť konfigurácie *nie je*: „Fotón z B smerom do E, a fotón z C smerom do F.“ Potom by výsledné konfigurácie v prípade 1 a prípade 4 neboli rovnaké. Prípad 1 by bol „Fotón z B do E, fotón z C do F“ a prípad 4 by bol „Fotón z B do F, fotón z C do E.“ *Keby* konfigurácie sledovali trasy fotónov, toto by boli dve rôzne konfigurácie.

Nemohli by sme teda sčítať tieto dve amplitúdy, a tie by sa navzájom nezrušili. Mali by sme tieto amplitúdy v dvoch rôznych konfiguráciách. Výsledné amplitúdy by mali nenulové druhé mocniny absolútnych hodnôt. A keby sme zbehli tento experiment, zistili by sme (asi v polovici prípadov), že detektor 1 a detektor 2 zaznamenali každý po jednom fotóne. Čo sa nedeje, ak sú moje výpočty správne.

Konfigurácie si nesledujú, odkiaľ prišla ktorá častica. Totožnosť konfigurácie je iba „Fotón tu, fotón tam; elektrón tu, elektrón tam.“ Bez ohľadu na to, ako sa do tejto situácie dostanete, pokiaľ sú tam tie isté druhy častíc na tých istých miestach, počíta sa to ako tá istá konfigurácia.

Opakujem, že otázka: „Aké informácie zahŕňa štruktúra konfigurácie?“ má *experimentálne dôsledky*. Podľa experimentu dokážete vydedukovať, ako samotná skutočnosť zaobchádza s konfiguráciami.

V klasickom vesmíre by neboli žiadne experimentálne dôsledky. Keby bol fotón ako biliardová guľa, ktorá išla buď jedným smerom alebo druhým, a keby konfigurácie boli naše názory na možné stavy, v akých môže byť systém, a keby sme namiesto amplitúd mali pravdepodobnosti, nebol by v tom rozdiel, či by sme sledovali pôvod fotónov alebo túto informáciu odhadzovali.

V klasickom vesmíre by som mohol priradiť pravdepodobnosť 25 % tomu, že oba fotóny pôjdu do E, 25 % pravdepodobnosť, že oba fotóny pôjdu do F, 25 % pravdepodobnosť, že fotón z B pôjde do E a fotón z C pôjde do F, a 25 % pravdepodobnosť, že fotón z B pôjde do F a fotón z C pôjde do E. Alebo, keďže mi je *osobne* jedno, ktorý z posledných dvoch prípadov nastane, by som sa mohol rozhodnúť zlúčiť tieto dve možnosti do jednej možnosti, sčítať ich pravdepodobnosti, a povedať iba: „pravdepodobnosť 50 %, že každý detektor dostane po jednom fotóne.“

Pri pravdepodobnostiach môžeme spájať udalosti dokopy, ako sa nám zachce – kresliť svoje hranice okolo množín možných svetov, ako sa nám zachce – a čísla stále vyjdú rovnaké.

Nemôžete však vo svojom modeli svojvoľne zlučovať konfigurácie, ani ich rozdeľovať, lebo nedostanete rovnaké experimentálne predpovede. Náš čarovný nástroj nám povie pomery druhých mocnín absolútnych hodnôt. Keď sčítate dve komplexné čísla, druhá mocnina absolútnej hodnoty súčtu nie je súčet druhých mocnín absolútnych hodnôt sčítancov:

$$|C1 + C2|^2 != |C1|^2 + |C2|^2$$

Napríklad:

$$|(2 + i) + (1 + -i)|^2 = |(3 + 0i)|^2 = 3^2 + 0^2 = 9$$

$$|(2 + i)|^2 + |(1 + -i)|^2 = (2^2 + 1^2) + (1^2 + 1^2) = (4 + 1) + (1 + 1) = 7$$

Alebo, v reči nášho dnešného experimentu, prúdy (1 + 0i) a (-1 + 0i) sa navzájom zrušili, ich súčtom je 0, a jej druhá mocnina je 0, zatiaľ čo druhé mocniny absolútnych hodnôt sčítancov by boli 1 a 1.

Keby namiesto druhej mocniny absolútnej hodnoty bola naším čarovným nástrojom nejaká lineárna funkcia – hocijaká funkcia, kde F(X + Y) = F(X) + F(Y) – potom by všetka kvantovitosť okamžite zmizla a nahradila by ju klasická fyzika. (*Iná* klasická fyzika, nie rovnaká ilúzia klasickosti, akú si halucinujeme zvnútra vyšších úrovní organizácie v našom kvantovom svete.)

Keby amplitúdy boli iba pravdepodobnosti, nemohli by sa navzájom zrušiť, keď sa ich toky zrazia. Keby konfigurácie boli iba stavy poznania, mohli by ste ich preusporiadať, ako by sa vám zachcelo.

Ale konfigurácie sú priklincované na mieste, nerozdeliteľné a nespojiteľné bez zmeny zákonov fyziky.

A časť toho, čo je priklincované, je spôsob, ako konfigurácie zaobchádzajú s viacerými časticami. Konfigurácia hovorí „fotón tu, fotón tam“, nie „*tento* fotón tu, *tamten* fotón tam“. „*Tento* fotón tu, *tamten* fotón tam“ nemá inú totožnosť ako „*tamten* fotón tu, *tento* fotón tam“.

Výsledok, viditeľný z dnešného experimentu, je, že si nemôžete faktorizovať fyziku nášho vesmíru, aby bola o časticiach s individuálnymi identitami.

Časť dôvodu, prečo majú ľudia problém vyrovnať sa s *dokonale normálnou* kvantovou fyzikou je, že ľudia sa bizarne snažia faktorizovať skutočnosť na súčet jednotlivo skutočných biliardových gulí.

Ha ha! Hlúpi ľudia.

*→

232. Rôzne konfigurácie

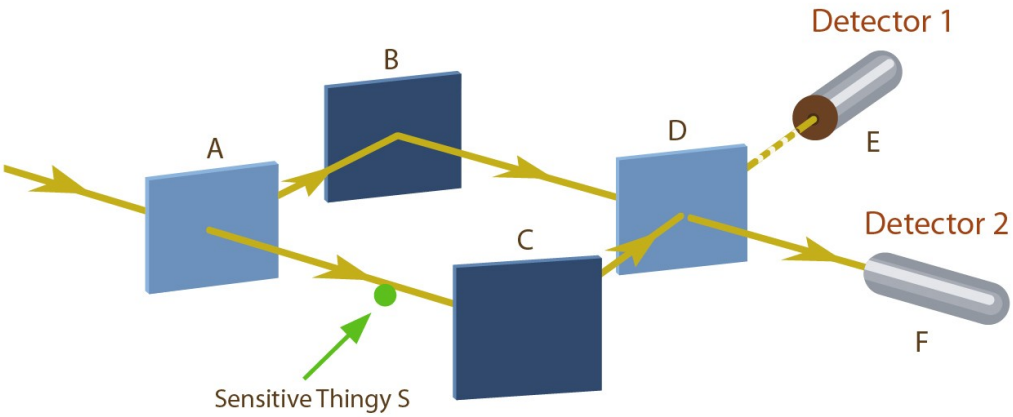
Pokus v predchádzajúcej kapitole priniesol dve hlavné ponaučenia:

Po prvé sme videli, že pretože toky amplitúdy sa môžu navzájom zrušiť, a pretože naša čarovná miera, druhá mocnina absolútnej hodnoty, nie je lineárna, totožnosť konfigurácií je pevne daná – nemôžete si preorganizovať konfigurácie tak, ako si môžete preskupiť možné svety. Ktoré konfigurácie sú totožné, a ktoré sú rôzne, má experimentálne dôsledky; je to pozorovateľný fakt.

Po druhé sme videli, že konfigurácie sa týkajú viacerých častíc. Ak do prístroja vchádzajú dve častice, neznamená to, že sú dve počiatočné konfigurácie. Namiesto toho je totožnosť počiatočnej konfigurácie „dva fotóny vchádzajú“. (V ideálnom prípade by každá konfigurácia, o ktorej hovoríme, zahŕňala všetky častice v experimente – vrátane častíc, z ktorých sa skladajú zrkadlá a detektory. A v skutočnom vesmíre sa každá konfigurácia týka *všetkých* častíc... *všade*.)

To, čo tvorí rôzne konfigurácie, nie sú rôzne častice. Každá konfigurácia sa týka všetkých častíc. To, čo robí konfigurácie rôznymi, sú častice nachádzajúce sa v rôznych polohách – aspoň jedna častica v inom stave.

Urobíme si dôležitú ukážku...



Obrázok 232.1

Na obrázku 232.1 je rovnaký pokus ako na obrázku 230.2, s jednou dôležitou zmenou: Medzi A a C sme umiestnili citlivú večičku S. Dôležitou vlastnosťou S je, že keď okolo S prejde fotón, S skončí v trochu inom stave.

Povedzme, že dva možné stavy S sú *Áno* a *Nie*. S začína v stave *Nie*, a skončí v stave *Áno*, ak okolo prejde fotón.

Počiatočná konfigurácia je potom:

„Fotón smerom do A; S v stave *Nie*.“ (-1 + 0i)

→ http://lesswrong.com/lw/pe/joint_configurations/

Nasleduje akcia polopriepustného zrkadla v A. V predchádzajúcej verzii tohto pokusu, bez tej citlivej vecičky, boli dve výsledné konfigurácie „z A do B“ s amplitúdou -i, a „z A do C“ s amplitúdou -1. Teraz sme však do systému zaviedli nový prvok a všetky konfigurácie sa týkajú všetkých častíc, takže každá konfigurácia spomína tento nový prvok. Toky amplitúdy z počiatočnej konfigurácie sú teda:

„Fotón z A do B; S v stave *Nie*.“ (0 + -i)

„Fotón z A do C; S v stave *Áno*.“ (-1 + 0i)

Ďalej, akcia úplných zrkadiel v B a C:

„Fotón z B do D; S v stave *Nie*.“ (1 + 0i) „Fotón z C do D; S v stave *Áno*.“ (0 + -i)

A potom akcia polopriepustného zrkadla v D, na amplitúdach pritekajúcich z oboch vyššie uvedených konfigurácií:

1. „Fotón z D do E; S v stave *Nie*.“ (0 + i)
2. „Fotón z D do F; S v stave *Nie*.“ (1 + 0i)
3. „Fotón z D do E; S v stave *Áno*.“ (0 + -i)
4. „Fotón z D do F; S v stave *Áno*.“ (1 + 0i)

Keď sme tento experiment robili bez citlivej vecičky, toky amplitúdy (1) a (3), (0 + i) a (0 + -i) pre konfiguráciu „z D do E“ sa navzájom zrušili. Nezostala nám žiadna amplitúda pre fotón idúci do detektora 1 (vysoko hore na experimentálnej úrovni sme nikdy nevideli, že by fotón narazil do detektora 1).

Ale v tomto prípade sú toky amplitúdy (1) a (3) pre rôzne konfigurácie; prinajmenšom jedna vec, S, je v rôznom stave medzi (1) a (3). Amplitúdy sa navzájom nezrušia.

Keď zamávame naším čarovným detektorom pomeru druhých mocnín absolútnych hodnôt nad výslednými štyrmi konfiguráciami, zistíme, že sa ich druhé mocniny absolútnych hodnôt rovnajú: každá po 25 %. Vysoko hore na úrovni skutočného sveta zistíme, že fotón má rovnakú šancu naraziť do detektora 1 alebo detektora 2.

Všetko horeuvedené je pravda, dokonca aj keď sa my ako výskumníci nestaráme o stav S. Na rozdiel od možných svetov, konfigurácie nemožno svojvoľne preskupovať. Zákony *fyziky* hovoria, že tieto dve konfigurácie sú rôzne; nie je to otázka toho, ako si *my* môžeme čo najpohodlnejšie rozdeliť svet.

Všetko horeuvedené je pravda, dokonca aj keď sa neunúvame pozrieť na stav S. Konfigurácie (1) a (3) sú vo fyzike rôzne, dokonca aj keď o tomto rozdiele nevieme.

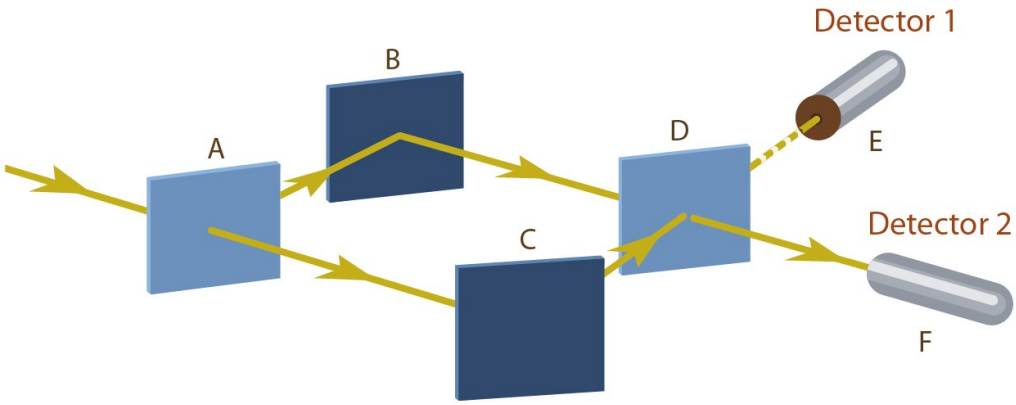
Všetko horeuvedené je pravda, dokonca aj keď nevieme, že S existuje. Konfigurácie (1) a (3) sú rôzne bez ohľadu na to, či *my* máme alebo nemáme pre tieto dve možnosti oddelené *myšlienkové reprezentácie*.

Všetko horeuvedené je pravda, dokonca aj keď sme vo vesmíre a S vyšle nový fotón smerom k medzihviezdnej prázdnote dvoma rôznymi smermi, podľa toho, či náš sledovaný fotón preletel okolo alebo nie. Takže by sme nemohli *nikdy* zistiť, či S bolo *Áno* alebo *Nie*. Stav S by bol stelesnený fotónom vyslaným do ničoty. Ten stratený fotón môže byť implikované neviditeľné, a stav S môže byť pragmaticky nezistiteľný; ale tie konfigurácie sú stále rôzne.

(Hlavný dôvod, prečo by toto mohlo *nefungovať*, je keby S bolo posunuté, ale keby malo v priestore konfigurácií pôvodný rozptyl väčší než toto posunutie. Potom by ste sa nemohli spoliehať, že toto posunutie rozdelí distribúciu amplitúdy v priestore konfigurácií na dve oddelené kôpky. V skutočnosti sa toto všetko odohráva v diferencovateľnej distribúcii amplitúdy v spojitom priestore konfigurácií.)

Konfigurácie nie sú názory. Ich rozdielnosť je objektívny fakt s experimentálnym dôsledkami. Konfigurácie sú rôzne aj keď nikto nevie o stave S; rôzne aj keď to nikdy žiaden inteligentný tvor nezistí. Konfigurácie sú rôzne, dokiaľ je aspoň *jedna* častica *hocikde* vo vesmíre v inej polohe. To sa dá experimentálne dokázať.

Prečo to zdôrazňujem? Pretože dávno za temných čias, keď nikto nerozumel kvantovej fyzike...



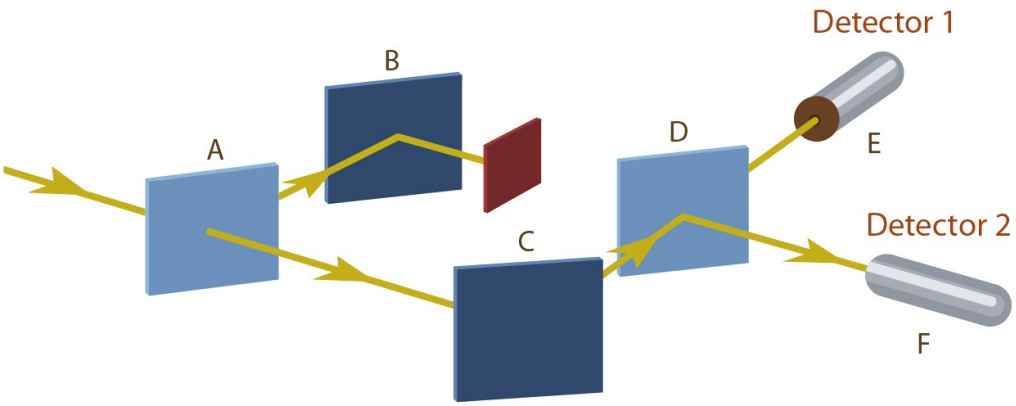
Obrázok 232.2

Dobre, predstavte si, že nemáte ani tušenie, čo sa v skutočnosti odohráva, a vyskúšate experiment na obrázku 2, a v detektore 1 sa neobjaví žiaden fotón. Paráda.

Tiež objavíte, že keď dáte prekážku medzi B a D, *alebo* prekážku medzi A a C, fotóny sa objavia pri detektore 1 a detektore 2 v rovnakom pomere. Ale vždy iba pri jednom – cvakne detektor 1 alebo detektor 2, nie oba naraz.

Takže áno, *vyzerá* to, ako keby ste mali do činenia s časticou – ten fotón je v jednom čase iba na jednom mieste, vždy keď ho *vy* vidíte.

A napriek tomu je tu akýsi... *tajomný jav*... ktorý nedovoľuje fotónu objaviť sa v detektore 1. A tento tajomný jav závisí od toho, či je *možné*, aby fotón šiel oboma smermi. Dokonca aj keď sa fotón objaví iba v jednom alebo iba v druhom detektore, čo *podľa vášho názoru* ukazuje, že fotón je zároveň iba na jednom mieste.



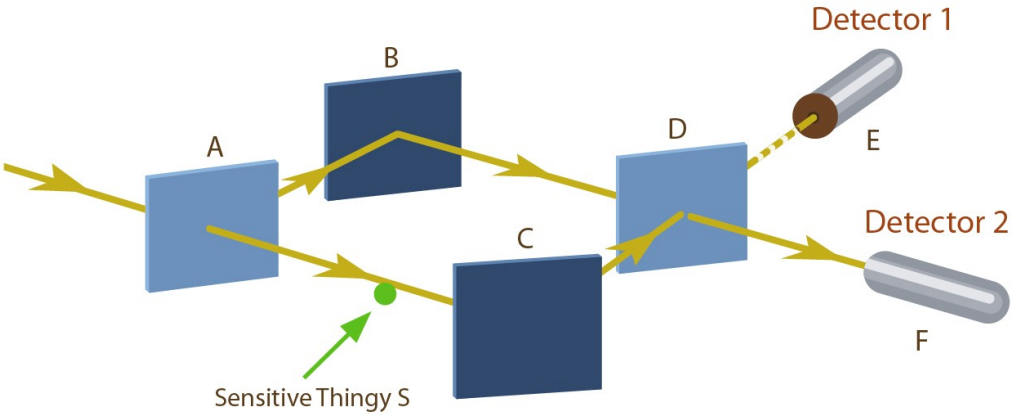
Obrázok 232.3

Vďaka čomu celý tento vzor experimentov vyzerá veľmi čudesne! Napokon, ten fotón buď ide z A do C, alebo z A do B; jedno alebo druhé. (Teda, to by ste si mysleli, keby ste sa inštinktívne snažili rozbiť skutočnosť na jednotlivo skutočné častice.) Ale keď zablokujete jednu alebo druhú trasu, začnete dostávať iné experimentálne výsledky!

Je to akoby mal fotón *dovolené* ísť oboma smermi, hoci (mysleli by ste si) ide iba jedným alebo druhým smerom. A dokáže to *zistiť*, keď sa ho pokúšate zablokovať, aj keď *tam* v skutočnosti nejde – že keby bol išiel tým smerom, bol by narazil na prekážku, a nebol by narazil na žiaden detektor.

Je to akoby púhe *možnosti* mali kauzálne účinky, v rozpore s tým, čo si myslíme, že slovo „skutočný“ zvyčajne *znamená*...

Ale je trochu skoro skákať k *takýmto* záverom, keď nemáte úplný obraz toho, čo sa odohráva vnútri experimentu.



Obrázok 232.4

Tak vám napadne umiestniť senzor medzi A a C, ako na obrázku 232.4, aby ste vedeli povedať, ktorým smerom ten fotón v ktorom prípade *naozaj* išiel.

A vtedy tajomný jav zmizne.

Akože, aké šialené je toto? Aký druh paranoje to vyvolá v nejakom chudákovi vedcovi?

Dobre, takže v 21. storočí sme si uvedomili, že aby sme „poznali“ históriu fotónu, častice, z ktorých sa skladá váš mozog, musia byť korelované s históriou toho fotónu. Ak mať maličkú citlivú vecičku S, ktorá koreluje s históriou fotónu, stačí na rozlíšenie konečných konfigurácií a zabráni tokom amplitúdy vo vzájomnom zrušení sa; potom celý senzor s digitálnym displejom, a tobôž ľudský mozog, umiestnia *kvadrilióny* častíc do iných pozícií a zabránia tokom amplitúdy, aby sa navzájom zrušili.

Ale keby ste na toto zatiaľ nedošli...

Potom by ste sa zamýšľali nad tým, ako ten senzor zahnal onen Tajomný Jav a mysleli by ste si:

Ten fotón nechce len mať *fyzicky* umožnené ísť ľubovoľným smerom. Nie je to malá vlna idúca pozdĺž nezablokovanej cesty, pretože potom by stačilo, že táto cesta nie je fyzicky zablokovaná.

Nie... nechce mi to dovoliť *vedieť*, ktorým smerom ten fotón šiel.

Ten tajomný jav... *nechce, aby som sa naň zblízka pozeral*... zatiaľ čo robí tú svoju tajomnú vec.

Nie sú to *fyzikálne možnosti*, ktoré majú účinok na skutočnosť... iba *epistemické možnosti*. Ak *viem*, ktorým smerom ten fotón šiel, už nie je *uveriteľné*, že išiel druhým smerom... čo odstrihne tento tajomný jav rovnako účinne ako umiestnenie prekážky medzi B a D.

Musím sa *zdržať pozorovania*, ktorým smerom ten fotón šiel, aby vždy skončil v detektore 2. Musí byť *uveriteľné*, že ten fotón mohol ísť aj do B, aj do C. Rozhodujúcim faktorom je to, čo môžem *vedieť*, bez ohľadu na to, ktoré fyzické cesty som nechal otvorené alebo zatvorené.

ZASTAVTE TLAČIARNE! MYSEĽ JE PREDSA LEN ZÁKLADOM! VEDOMIE URČUJE NAŠE POKUSNÉ VÝSLEDKY!

Takéto veci si *stále môžete prečítať*. V *učebniciach fyziky*. Dokonca aj dnes, keď už to väčšina teoretických fyzikov vie lepšie. Zastavte tlačiarne. Prosím, zastavte tlačiarne.

Po bitke je každý generálom; a je také ľahké v spätnom pohľade povedať, že existovali isté náznaky, že táto interpretácia nie je správna.

Napríklad, ak postavíte senzor medzi A a C, *ale neprečítate ho*, ten tajomný jav *aj tak* zmizne, a ten fotón aj tak niekedy skončí v detektore 1. (Aha, lenže ste si to *mohli* prečítať, a možnosti sú teraz skutočné...)

Ale nemusí to byť ani *senzor*, vami postavený vedecký prístroj. Jediná častica, ktorú niečo dostatočne odstrčí, preruší interferenciu. Stačí na to fotón, ktorý odletí preč niekam, kde ho nikdy viac neuvidíte. Tam nie je veľa ľudského zásahu. Tam nie je veľa vedomia.

Možno ste pred spustením dualistického požiarneho poplachu o fyzikálnej výnimočnosti ľudských mozgov mohli experimentálne overiť, či kameň nedokáže zohrať rovnakú rolu v zaháňaní toho Tajomného Javu ako ľudský výskumník?

Ale to je spätný pohľad, a je ľahké rozhodovať sa v spätnom pohľade. *Naozaj* si myslíte, že by ste konali lepšie než John von Neumann, keby ste vtedy žili? Pointou tohto druhu spätnej analýzy je opýtať sa, aké celkom všeobecné náznaky ste si mohli všimnúť, a či existujú nejaké podobné náznaky, ktoré ignorujete ohľadom terajších tajomstiev.

Ale *je* to trochu zahanbujúce, že aj *po* vypracovaní teórie amplitúd a konfigurácií – keď už tá teória dávala jednoznačnú predpoveď, že by na to stačila ľubovoľná posunutá častica – raným vedcom to *stále* nedošlo.

Lebo, ako vidíte... prijalo sa ako Všeobecne Známe, že konfigurácie sú možnosti, že záleží na epistemickej možnosti, že amplitúdy sú veľmi čudným druhom čiastočnej informácie, a že vedomé pozorovanie zaháňa kvantovitosť. A že je najlepšie o celej tejto veci príliš tvrdo nerozmýšľať, dokiaľ vám experimentálne predpovede vychádzajú správne.

*→

233. Predpoklad kolapsu

Makroskopická dekoherencia - známa aj ako „mnohé svety“ - je myšlienka, že známe kvantové zákony, ktoré riadia mikroskopické javy, jednoducho riadia všetky úrovne bez zmeny. Keď ešte ľudia nevedeli o dekoherencii - skôr než niekomu napadlo, že zákony tak presne odvolené pre mikroskopickú fyziku by mohli platiť univerzálne na všetkých úrovniach - čo si ľudia *mysleli*, že sa deje?

Počiatočné uvažovanie išlo asi nejako takto, z:

> „Keď moje výpočty ukázali amplitúdu -(1/3)i pohltenia tohto fotónu, moja experimentálna štatistika ukázala, že fotón bol pohltený zhruba v 107 prípadoch z 1000, čo dobre zodpovedá 1/9, druhej mocnine absolútnej hodnoty.“

cez

> „Amplitúda *je* pravdepodobnosť (keď z nej urobíme druhú mocninu absolútnej hodnoty).“

do:

> „Keď niečo nameriate a *viete, že sa to nestalo*, táto *pravdepodobnosť* sa zmení na nulu.“

Pri doslovnom čítaní z toho vyplýva, že samotný poznatok – alebo dokonca vedomie – spôsobuje kolaps. Čo bola naozaj podoba teórie, ktorú predložil Werner Heisenberg!

Ale ľudia začali byť čoraz nervóznejší z predstavy zavlečenia dualistického jazyka do základov fyziky – a veru mali byť prečo! A tak sa pôvodná verzia nahradila predstavou objektívneho „kolapsu“, ktorý zničí všetky časti vlnovej funkcie okrem jednej, a ktorý nastane niekedy predtým než superpozícia narastie na ľuďmi pozorovateľnú úroveň.

Nuž, keď už predpokladáte, že časti vlnovej funkcie len tak zmiznú, môžete si položiť otázku:

> „Prežije iba *jedna* časť? Možno existujú mnohé svety, ktoré prežijú, ale prežijú s frekvenciou určenou integrálom druhej mocniny ich absolútnej hodnoty, takže typický prežívajúci svet štatisticky zodpovedá Bornovmu pravidlu.“

Napriek tomu teórie kolapsu zvažované modernou akadémiou predpokladajú, že prežije iba *jeden* svet. Prečo?

Teórie kolapsu boli vytvorené v čase, keď žiadnemu fyzikovi *jednoducho nenapadlo*, že by *mohol* existovať viac než jeden svet! Ľudia brali ako samozrejmosť, že meranie má jeden výsledok – bol to predpoklada taký hlboký, až bol neviditeľný, pretože to bolo to, čo *videli, že sa deje*. Teórie kolapsu boli vytvorená, aby vysvetlili, *prečo má meranie iba jeden výsledok*, namiesto toho, (v úplnej všeobecnosti) *prečo experimentálne štatistiky zodpovedajú Bornovmu pravidlu*.

Z podobných dôvodov „predpoklady kolapsu“ zvažované akademicky predpokladajú, že kolaps nastane *skôr*, než sa nejaký človek dostane do superpozície. Experimenty však postupne vyvracajú

→ http://lesswrong.com/lw/pf/distinct_configurations/

možnosť „kolapsu“ v čoraz väčších previazaných systémoch. Zdá sa, že sa chystá experiment, ktorý ukáže kvantovú superpozíciu na škále 50 mikrometrov, čo je viac než väčšina neurónov, a blíži sa to k priemeru niektorých ľudských vlasov!

Prečo teda niekto v tejto hre neskočí dopredu a neopýta sa:

> „Pozrite, stále posúvame tento predpokladaný kolaps na neskôr a neskôr. Čo ak ten kolaps nastáva až keď superpozícia dosiahne veľkosť planéty a nastane podstatná odchýlka – povedzme, čo ak vlnová funkcie Zeme kolabuje zhruba raz za minútu? Potom, aj keď si väčšina prežívajúcich Zemí v ľubovoľnom čase *pamätá* dlhú históriu kvantových experimentov, ktoré zodpovedajú Bornovým štatistikám, prevažná väčšina týchto Zemí začne dostávať nie-bornovské výsledky kvantových experimentov, a potom o minútu neskôr náhle prestane existovať.“

Prečo nemajú *takéto* teórie kolapsu veľa nasledovníkov v akadémii, medzi tým množstvom ľudí, ktorí si zrejme myslia, že je okej, keď časť vlnovej funkcie len tak zmizne? Najmä ak experimenty dokazujú superpozíciu v čoraz väčších systémoch?

Cynik by mohol navrhnúť, že dôvodom pretrvávajúcej podpory pre kolaps nie je *fyzikálna uveriteľnosť*, že by veľké časti vlnovej funkcie náhle mizli, ani nádej, že sa nejako vysvetlia Bornove štatistiky. Pointa je udržať si intuitívnu príťažlivosť tvrdenia „nepamätám si, že by merania mali viac než jeden výsledok, preto sa stala iba jedna vec; nepamätám si rozdvojenie, preto musí existovať iba jedno ja“. Nepamätáte si vlastný zánik, takže ľudia v superpozícii nemôžu nikdy kolabovať. Teória, ktorá by sa opovážila odporovať intuícii, by nepochopila celú pointu. To už by ste rovnako dobre mohli prejsť k dekoherencii.

To by mohol navrhnúť cynik.

Ale iste je príliš zavčasu útočiť na motívy zástancov kolapsu. Toto je púhy argument ad hominem. A čo skutočná fyzikálna dôveryhodnosť teórií kolapsu?

Nuž, po prvé: Má teória kolapsu nejaké experimentálne potvrdenie? Nie.

Keď sme toto vyriešili...

Keby kolaps naozaj fungoval tak, ako jeho zástancovia tvrdia, že funguje, bol by to:

1. Jediný nelineárny vývoj v celej kvantovej mechanike.
2. Jediný neunitárny vývoj v celej kvantovej mechanike.
3. Jediný nediferencovateľný (vlastne nespojitý) jav v celej kvantovej mechanike.
4. Jediný jav v celej kvantovej mechanike, ktorý je nelokálny v konfiguračnom priestore.
5. Jediný jav v celej fyzike, ktorý porušuje súmernosť CPT.
6. Jediný jav v celej fyzike, ktorý porušuje Liouvilleovu vetu (má mapovanie viacerých počiatočných stavov na jeden výsledok).
7. Jediný jav v celej fyzike, ktorý je nekauzálny / nedeterministický / skutočne náhodný.
8. Jediný jav v celej fyzike, ktorý je nelokálny v časopriestore a šíri účinok rýchlejšie ako svetlo.

ČO EŠTE MUSÍ TEN PREKLIATY PREDPOKLAD KOLAPSU *UROBIŤ*, ABY HO FYZICI ODVRHLI? ZABIŤ PREKLIATE ŠTENIATKO?

*→

234. Dekoherencia je jednoduchá

Epištola fyzikom:

Keď som bol malým chlapcom, môj otec, fyzik s PhD, ma prísne varoval pred miešaním sa do záležitostí fyzikov; povedal, že je beznádejné pokúšať sa porozumieť fyzike bez formálnej matematiky. Bodka. Žiadne výnimky. Čítal som však vo Feynmanových populárnych knihách, že ak

→ http://lesswrong.com/lw/q6/collapse_postulates/

fyzike naozaj rozumiete, mali by ste ju dokázať vysvetliť nefyzikovi. Veril som Feynmanovi viac než môjmu otcovi, pretože Feynman vyhral Nobelovu cenu, a môj otec nie.

Až neskôr – keď som si naozaj čítal *Feynmanove Lekcie* – som si uvedomil, že mi otec povedal jednoduchú a úprimnú pravdu. Žiadna matematika = žiadna fyzika.

Ja som povolaním bayesovec, nie fyzik. Hoci som bol vychovaný, aby som sa nemiešal do záležitostí fyzikov, bol som k tomu donútený občasným hrubým zneužívaním troch pojmov: *jednoduché, falzifikovateľné*, a *testovateľné*.

Predchádzajúci úvod je nato, aby ste sa nesmiali a nehovorili: „Samozrejme, že viem, čo tieto slová znamenajú!“ Je v tom matematika. Nasledovať bude zopakovanie bodov z kapitoly Viera v implicitné neviditeľné, ako sa týkajú kvantovej fyziky.

Začnime poznámkou, ktorá ma poslala na celú túto cestu, ktorú som videl vo viacerých verziách; parafrázovaná znie:

> Interpretácia mnohých svetov v kvantovej mechanike predpokladá, že existuje obrovské množstvo iných svetov existujúcich paralelne s naším. Occamova britva hovorí, že by sme nemali zbytočne pridávať veci.

Pri všetkej férovosti však treba dodať, že tí, čo to hovoria, väčšinou aj priznajú:

> Ale toto nie nie všeobecne prijímaná aplikácia Occamovej britvy; niekorí hovoria, že Occamova britva sa vzťahuje na zákony, ktorými sa model riadi, nie na počet predmetov vnútri modelu.

Je teda dobré, že všetci uznávame opačné argumenty a hovoríme obe strany tohto príbehu...

Ale predstavte si, že by ste mali *vypočítať* jednoduchosť teórie.

Pôvodná formulácia Williama z Ockhamu znela:

Lex parsimoniae: Entia non sunt multiplicanda praeter necessitatem.

„Zákon šetrnosti: Veci by sa nemali množiť viac než treba.“

Ale toto je kvalitatívna rada. Nestačí na povedanie, či jedna teória vyzerá jednoduchšie alebo zložitejšie než iná – musíte priradiť nejaké číslo; a toto číslo musí byť zmysluplné, nemôžete si ho len tak vymyslieť. Prekročiť túto priepasť je ako rozdiel medzi tým, či viete odhadom povedať, ktoré veci sa pohybujú „rýchlo“ a „pomaly“, a keď začnete merať a počítať rýchlosti.

Predstavte si, že by ste skúsili povedať: „Spočítajte slová – taká komplikovaná je daná teória.“

Robert Heinlen raz tvrdil (dúfam, že ironicky), že „najjednoduchšie vysvetlenie“ je vždy: „Pani z dolného konca ulice je bosorka; ona to urobila.“. Desať slov – máloktorý fyzikálny článok dokáže tromfnúť toto.

Zoči-voči tejto výzve máte na výber dve cesty.

Po prvé, môžete sa opýtať: „Tá žena z dolného konca ulice je *čo*?“ Len preto, že má angličtina pre nejaký pojem jedno slovo, to neznamená, že samotný tento pojem je jednoduchý. Predstavte si, že sa rozprávate s mimozemšťanom, ktorý nevie nič o bosorkách, ženách, ani uliciach – ako dlho by vám trvalo vysvetliť mu vašu teóriu? Ešte lepšie, predstavte si, že máte napísať počítačový program stelesňujúci vašu hypotézu a vypisujúci to, čo považujete za predpovede vašej hypotézy – ako veľký by musel byť tento počítačový program? Povedzme, že je vašou úlohou predpovedať časovú postupnosť meraných polôh skaly, ktorá sa kotúľa dole kopcom – ak napíšete podprogram, ktorý simuluje bosorky, asi vám nepomôže upresniť, kadiaľ sa skala bude kotúľať – tento podprogram navyše iba predlžuje váš kód. Mohli by ste však zistiť, že váš kód nevyhnutne zahŕňa podprogram umocňujúci čísla na druhú.

Po druhé, môžete sa opýtať: „Tá žena z dolného konca ulice je bosorka; urobila *čo*?“ Predpokladajme, že chcete opísať nejakú udalosť tak presne, ako sa len dá pri dostupných indíciách – opäť, povedzme postupnosť vzdialenosť/čas skaly kotúľajúcej sa dole kopcom. Môžete svoje vysvetlenie uviesť slovami: „Tá žena z dolného konca ulice je bosorka“, ale potom váš kamarát povie: „*Čo* urobila?“ a vy odpoviete: „Spôsobila, že sa kameň po prvej sekunde skotúľal o jeden meter, po tretej sekunde o deväť metrov...“ Dať na úvod správy: „Tá žena z dolného konca ulice je bosorka“ vám nepomôže *komprimovať* zvyšok vášho popisu. Celkovo nakoniec pošlete dlhšiu správu než bolo potrebné – dáva

väčší zmysel vynechať ten úvod s „bosorkou“. Na druhej strane, ak venujete chvíľu času Galileovi, môže vám to pomôcť značne skomprimovať nasledujúcich päť tisíc podrobných časových postupností pre skaly kotúľajúce sa dole kopcom.

Ak pôjdete tou prvou cestou, skončíte s tým, čo sa nazýva Kolmogorova zložitosť a Solomonoffova indukcia. Ak pôjdete tou druhou cestou, skončíte s tým, čo sa nazýva minimálna dĺžka správy.

> Ach, takže si môžem vybrať medzi rôznymi definíciami jednoduchosti?

Nie, v skutočnosti sa dokázalo, že oba tieto formalizmy vo svojej najrozvinutejšej podobe sú ekvivalentné.

> A predpokladám, že mi teraz povieš, že oba formalizmy sú na strane „Occam znamená počítať pravidlá, nie počítať predmety“.

Viacmenej. Pri minimálnej dĺžke správy pokiaľ viete svojmu kamarátovi povedať presný recept, ktorým sa môže v myšlienkach riadiť, aby získal časovú postupnosť padajúcej skaly, nestaráme sa o to, koľko myšlienkovej námahy ho stojí riadiť sa týmto receptom. Pri Solomonoffovej indukcii počítame bity v kóde programu, nie množstvo pamäte, ktoré program pri spustení zaberie. „Veci“ sú tu riadky kódu, nie simulované predmety. A ako som povedal, v konečnom dôsledku tú tieto dva formalizmy ekvivalentné.

Teraz skôr než sa pustím do ďalších podrobností ohľadom formálnej jednoduchosti, dovoľte mi odbočiť a venovať sa tejto námietke:

> No a čo? Prečo si nemôžem vymyslieť svoj *vlastný* formalizmus, ktorý robí veci inak? Prečo by som mal venovať pozornosť spôsobu, ktorým si sa ty rozhodol robiť veci, v tvojej oblasti? Máš nejaké *experimentálne* indície, ktoré ukazujú, že by som mal robiť veci po tvojom?

V skutočnosti áno, či tomu veríte alebo nie. Ale začnime na začiatku.

Pravidlo konjunkcie v teórii pravdepodobnosti hovorí:

$$P(X \wedge Y) \leq P(X)$$

Pre ľubovoľné výroky X a Y, pravdepodobnosť, že „X je pravda, a Y je pravda“ je *menšia alebo rovná* pravdepodobnosti, že „X je pravda (bez ohľadu na to, či Y je alebo nie je pravda)“. (Ak vám toto tvrdenie neznie príliš hlboké, dovoľte mi ubezpečiť vás, že je ľahké nájsť prípady, keď ľudský odhad pravdepodobnosti porušuje toto pravidlo.)

Zvyčajne nemôžete pravidlo $P(X \wedge Y) \leq P(X)$ aplikovať priamo na konflikt dvoch navzájom sa vylučujúcich hypotéz. Pravidlo konjunkcie sa vzťahuje priamo iba na tie prípady, kde z ľavej strany striktne vyplýva pravá strana. Navyše, táto konjunkcia je iba nerovnosť; nedáva nám ten druh kvantitatívnych výpočtov, ktoré chceme.

Pravidlo konjunkcie nám však dáva pravidlo monotónneho znižovania pravdepodobnosti: ako do príbehu pridávate viac podrobností, a každá dodatočná podrobnosť môže byť pravda alebo nepravda, pravdepodobnosť celého príbehu sa monotónne znižuje. Myslite na pravdepodobnosť ako na hodnotu, ktorá sa zachováva: máme k dispozícii iba obmedzené množstvo. Ako sa zvyšuje počet podrobností v príbehu, počet možných príbehov rastie exponenciálne, ale súčet ich pravdepodobností nemôže byť nikdy väčší než 1. Pre každý príbeh „X a Y“ existuje príbeh „X a ~Y“. Keď poviete *iba* príbeh „X“, dostanete *súčet* pre možnosti Y a ~Y.

A pridáte k X desať podrobností, z ktorých každá podrobnosť môže byť pravda alebo nepravda, potom tento príbeh musí súperiť s (2^10 – 1) ďalšími rovnako podrobnými príbehmi o vzácnu pravdepodobnosť. Ak na druhej strane stačí *iba* povedať X, môžete sčítať svoje pravdepodobnosti pre 2^10 príbehov

> ((X a Y a Z a …) alebo (X a ~Y a Z a …) alebo …).

Tie „veci“, ktoré počíta Occamova britva by mali jednotlivo stáť pravdepodobnosť; preto uprednostňujeme teórie, ktoré ich majú menej.

Predstavte si lotériu, ktorá predáva milión žrebov, kde je každý možný žreb predaný iba raz, a lotéria v čase žrebovania predala všetky žreby. Váš kamarát si kúpil jeden lístok za jeden dolár – čo

vám pripadá ako zlá investícia, lebo výhra je iba 500 000 dolárov. Váš kamarát však hovorí: „Ach, ale vezmi si alternatívne hypotézy: ‚Zajtra niekto vyhrá lotériu' a ‚Zajtra vyhrám lotériu'. Je jasné, že tá druhá hypotéza je podľa Occamovej britvy jednoduchšia; spomína iba jednu osobu a jeden žreb, zatiaľ čo tá prvá hypotéza je zložitejšia: spomína milión ľudí a milión žrebov!"

Povdať, že Occamova britva počíta iba zákony a nie predmety, nie je celkom správne: čo sa proti teórii počíta, to sú predmety, ktoré musí spomenúť *explicitne*, pretože to sú predmety, z ktorých nemožno urobiť iba *súčet*. Predstavte si, že vy a váš kamarát dumáte nad úžasným úderom v biliarde, kde vám povedia začiatočný stav biliardového stola a ktoré gule skončili v jamke, ale nie, ako ten úder vyzeral. Navrhnete teóriu, ktorá zahŕňa desať konkrétnych zrážok medzi desiatimi konkrétnymi guľami; váš priateľ konkuruje teóriou, ktorá zahŕňa päť konkrétnych zrážok medzi piatimi konkrétnymi guľami. Proti vašim teóriám sa počítajú *nielen* zákony, o ktorých tvrdíte, že riadia biliardové gule, ale aj všetky *konkrétne* biliardové gule, ktoré museli byť v nejakom *konkrétnom* stave, aby bola predpoveď vášho modelu úspešná.

Ak nameriate teplotu svojej obývačky ako 22 stupňov Celzia, nemá zmysel povedať: „Váš teplomer je asi pokazený; izba má s väčšou pravdepodobnosťou 20 stupňov. Lebo ak si zvážite všetky častice v tejto izbe, existuje exponenciálne viac stavov, v ktorých môžu byť, ak je teplota naozaj 22 stupňov – čo spôsobuje, že každý *konkrétny* stav je o to nepravdepodobnejší." Lebo bez ohľadu na to, ktorý presne zo stavov pri 22 stupňoch má vaša izba, môžete urobiť tú istú predpoveď (pre drvivú väčšinu týchto stavov), že váš teplomer nakoniec ukáže 22 stupňov, takže nie ste citliví na *presné* počiatočné podmienky. Nepotrebujete určiť presnú polohu všetkých molekúl vzduchu v izbe, takže toto sa nepočíta proti pravdepodobnosti vášho vysvetlenia.

Na druhej strane – vráťme sa k prípadu lotérie – predstavte si, že váš priateľ vyhrá desať lotérií po sebe. V tomto bode by ste mali začať podozrievať nejaký podvod. Hypotéza: „Môj kamarát vyhrá všetky lotérie" je omnoho zložitejšia než hypotéza „Každú lotériu niekto vyhrá". Ale tá prvá hypotéza predpovedá údaje omnoho presnejšie.

Vo formalizme minimálnej dĺžky správy, povedať: „Existuje jedna osoba, ktorá vždy vyhrá túto lotériu" na začiatku vašej správy zmenšuje váš popis toho, kto vyhral ďalších desať lotérií; stačí len dodať: „A tá osoba je Fred Smith" a správa je hotová. Porovnajte s: „Prvú lotériu vyhral Fred Smith, druhú lotériu vyhral Fred Smith, tretiu lotériu vyhral..."

Vo formalizme Solomonoffovej indukcie, pôvodná pravdepodobnosť „Môj kamarát vyhrá všetky lotérie" je nízka, pretože program, ktorý opisuje lotériu, teraz potrebuje explicitný kód na označenie vášho kamaráta; ale keďže tento program dokáže vytvoriť *hustejšiu distribúciu pravdepodobnosti* ohľadom potenciálnych víťazov lotérie než „Každú lotériu niekto vyhrá", môže podľa Bayesovho pravidla prekonať svoju pôvodnú nepravdepodobnosť a vyhrať ako hypotéza.

Ľubovoľná formálna teória Occamovej britvy by mala kvantitatívne definovať nielen „veci" a „jednoduchosť" ale aj „potrebné".

Minimálna dĺžka správy definuje potrebné ako „to, čo komprimuje správu".

Solomonoffova inducia priraďuje pôvodnú pravdepodobnosť každému počítačovému programu, kde celá distribúcia pre všetky možné počítačové programy nemá súčet väčší než 1. To sa dá dosiahnuť použitím dvojkového kódu, kde žiaden platný program nie je prefixom iného platného počítačového programu („bezprefixový kód"), pretože obsahuje značku stop. Potom je pôvodná pravdepodobnosť ľubovoľného programu jednoducho 2^-L(P), kde L(P) je dĺžka programu P v bitoch.

Samotné P môže byť program, ktorý dostane (potenciálne nulový) reťazec bitov a vypíše podmienenú pravdepodobnosť, že *nasledujúci* bit bude 1; toto robí z P pravdepodobnostnú distribúciu pre všetky dvojkové postupnosti. Táto verzia Solomonoffovej indukcie, pre každý reťazec, nám dáva zmes výsledných pravdepodobností, v ktorej dominujú najkratšie programy, ktoré najpresnejšie predpovedajú daný reťazec. Súčet pre túto zmes nám dáva predpoveď nasledujúceho bitu.

Výsledok je, že si vyžaduje viac bayesovskej indície – viac úspešných predpovedí, alebo presnejšie predpovede – aby ste obhájili zložitejšie hypotézy. Ale dá sa to urobiť; bremeno pôvodnej nepravdepodobnosti nie je nekonečné. Ak hodíte mincou štyrikrát, a vždy na nej padne hlava, nedôjdete

hneď k záveru, že táto minca dáva iba hlavy; ale ak na tejto minci padne hlava dvadsaťkrát po sebe, mali by ste to vážne zvážiť. A čo hypotéza, že minca je vyrobená tak, že produkuje „HTTHTT...“ v opakujúcom sa cykle? To je bizarnejšie – ale po sto hodoch mince by to popieral iba blázon.

Štandardná chémia hovorí, že v grame plynu vodíka je šesťsto tisíc miliónov miliárd atómov vodíka. To je znepokojujúce tvrdenie, ale existuje isté množstvo indície, ktoré stačilo na to, aby presvedčilo fyzikov vo všeobecnosti a vás konkrétne, že toto tvrdenie je pravdivé.

Teraz si položte otázku, koľko indície by bolo treba, aby vás presvedčila o teórii so šesťsto tisíc miliónmi miliárd samostatne upresnených fyzikálnych zákonov.

Prečo pôvodná pravdepodobnosť programu vo formalizme Solomonoffovej indukcie nezahŕňa meranie, koľko pamäte program spotrebuje, alebo ako dlho celkovo beží?

Jednoduchá odpoveď je: „Pretože priestorové a časové zdroje spotrebované programom nie sú navzájom sa vylučujúce možnosti. Na rozdiel od špecifikácie programu, kde na ľubovoľnom konkrétnom mieste môžete mať buď 1 alebo 0.“

Ale ešte jednoduchšia odpoveď je: „Pretože, historicky povedané, takáto heuristika nefunguje.“

Occamova britva sa používala ako námietka proti tvrdeniu, že hmloviny sú v skutočnosti vzdialené galaxie – zdalo sa, že to ohromne znásobuje počet vecí vo vesmíre. *Všetky tie hviezdy!*

Počas ľudskej histórie sa vesmír znova a znova zväčšoval. Verzii Occamovej britvy, ktorá by v každom prípade označila väčší vesmír za *nepravdepodobnejší,* by sa v historickej skúsenosti ľudstva darilo omnoho horšie.

Toto je časť tej „experimentálnej indície“, na ktorú som predtým narážal. Aj keď môžete zdôvodňovať teórie jednoduchosti na matematických základoch, stále je žiadúce, aby naozaj fungovali v praxi. (Tá druhá časť „experimentálnej indície“ pochádza od štatistikov / informatikov / výskumníkov umelej inteligencie, ktorí testovali, ktoré definície „jednoduchosti“ im dovolia zostavovať počítačové programy, ktorým sa empiricky darí predpovedať budúce údaje na základe minulých údajov. Tu sa asi ako najproduktívnejšia paradigma ukázala minimálna dĺžka správy, pretože je to veľmi prispôsobivý spôsob, ako rozmýšľať o problémoch skutočného sveta.)

Predstavte si vesmírnu loď, ktorej štartu sa prizeráte s veľkými oslavami; zrýchľuje smerom preč od vás a čoskoro putuje rýchlosťou 0,9 rýchlosti svetla. Ak bude rozširovanie vesmíru pokračovať v súlade s tvrdeniami dnešnej kozmológie, mal by byť nejaký bod v budúcnosti, kde – podľa vášho modelu skutočnosti – nebudete očakávať, že by ste mohli interagovať s touto vesmírnou loďou čo len v princípe; vzhľadom na vás zašla za kozmologický horizont, a fotóny, ktoré z nej odchádzajú, nedokážu predbehnúť rozširovanie vesmíru.

Mali by ste teda veriť, že <u>táto vesmírna loď doslova fyzicky zmizla z vesmíru</u> v okamihu, keď prešla za kozmologický horizont relatívne voči vám?

Ak veríte, že Occamova britva počíta predmety v modeli, potom áno, mali by ste. Akonáhle táto vesmírna loď zašla za váš kozmologický horizont, model, v ktorom táto vesmírna loď okamžite zmizla, a model, v ktorom táto vesmírna loď pokračuje ďalej, dávajú nerozlíšiteľné predpovede; nemajú voči sebe výhodu žiadnej bayesovskej indície. Ale jeden model obsahuje omnoho menej „vecí“; nemusí hovoriť o všetkých kvarkoch a elektrónoch a poliach, z ktorých sa táto vesmírna loď skladá. Takže je jednoduchšie predpokladať, že vesmírna loď zmizne.

Prípadne môžete povedať: „Počas mnohých experimentov som si zovšeobecnil isté zákony, ktorými sa riadia pozorované častice. Táto vesmírna loď sa skladá z takýchto častíc. Použijúc tieto zákony môže usúdiť, že táto vesmírna loď bude pokračovať ďalej po tom, čo prekročí kozmologický horizont, inak by sa porušili zákony zachovania, ktoré vidím, že platia v každej skúmateľnej situácii. Aby som predpokladal, že kozmická loď zmizne, musel by som pridať nový zákon: ‚Veci miznú, akonáhle prekročia môj kozmologický horizont.‘“

Verzia dekoherencie (alias mnohých svetov) kvantovej fyziky tvrdí, že sa merania riadia rovnakými kvantovo mechanickými pravidlami ako všetky ostatné fyzikálne procesy. Ak tieto pravidlá použijeme na makroskopické objekty celkom rovnakým spôsobom, ako na mikroskopické, vyjdú nám pozorovatelia v

stave superpozície. Tu je možné položiť veľa otázok, napríklad: „Prečo sa potom nezdá, že všetky kvantové merania majú pravdepodobnosť 50/50, keď rôzne verzie nás uvidia oba výsledky?“

Avšak námietka, že dekoherencia porušuje Occamovu britvu kvôli znásobeniu predmetov v modeli, je jednoducho nesprávna.

Dekoherencia si nevyžaduje, aby vlnová funkcia mala nejaký komplikovaný presný pôvodný stav. Mnohé svety nešpecifikujú všetky svoje svety ručne, ale vytvárajú ich pomocou kompaktných zákonov kvantovej mechaniky. Počítačový program, ktorý priamo simuluje QM, aby robil experimentálne predpovede, by si na spustenie vyžadoval hromadu pamäte – ale simulovanie vlnovej funkcie je exponenciálne náročné v *každej* verzii QM! Dekoherencia je skrátka viac taká. *Mnohé* fyzikálne objavy v ľudskej histórii, od hviezd po galaxie, od atómov po kvantovú mechaniku, výrazne zvýšili domnelú záťaž počítača, ktorý považujeme za náš vesmír.

Mnoho svetov nie je ziliónkrát komplikovanejšie podľa počtu svetov, rovnako ako hypotéza atómov nie je ziliónkrát komplikovanejšia podľa počtu atómov. Pre každého, kto chápe Occamovu britvu kvantitatívne, pojem „komplikovaný“ jednoducho toto neznamená.

Ako v historickom prípade galaxií, je možné, že si ľudia poplietli svoj *šok* z predstavy takého veľkého vesmíru s pravdepodobnostnou penaltou, a volajú si Occamovu britvu na pomoc. Ale ak existujú pravdepodobnostné penalty pre dekoherenciu, *veľkosť implikovaného vesmíru* ako taká jednoznačne nie je ich zdrojom!

Predstava, že dekoherentné svety sú dodatočné veci, ktoré Occamova britva penalizuje, je jednoducho jasne pomýlená. Nie je tak trochu správna. Nie je to argument, ktorý je slabý, ale napriek tomu platný. Nie je to obhájiteľná pozícia, ktorá sa dá podoprieť ďalšími argumentmi. Je to ako pravdepodobnostná teória úplne zle. Nedá sa to opraviť. Je to zlá matematika. 2 + 2 = 3.

⁂→

235. Dekoherencia je falzifikovateľná a testovateľná

Slová „falzifikovateľné“ a „testovateľné“ sa niekedy navzájom vymieňajú, a táto nepresnosť je cenou za rozprávanie po anglicky. V teórii pravdepodobnosti existujú dve rôzne vlastnosti, o ktorých tu chcem diskutovať, a jednu z nich budem nazývam „falzifikovateľné“ a tú druhú „testovateľné“, pretože sa mi zdá, že to tak najlepšie sedí.

Čo sa týka matematiky, tá začína, ako mnoho ďalších vecí, takto:

$$P(A_i|B) = P(B|A_i) \times P(A_i) / \text{Suma}_j\{ P(B|A_j) \times P(A_j) \}$$

Toto je Bayesova veta. Mám prinajmenšom dva rôzne kusy oblečenia, na ktorých je táto veta vytlačená, takže musí byť dôležitá.

Aby sme si to rýchlo pripomenuli, B tu znamená kus indície, Ai je nejaká hypotéza, o ktorej uvažujeme, a Aj sú konkurenčné, navzájom sa vylučujúce hypotézy. P(B | Ai) znamená „pravdepodobnosť, že uvidíme B, ak platí hypotéza Ai“ a P(Ai | B) znamená „pravdepodobnosť, že platí hypotéza Ai, ak vidíme B“.

Matematický jav, ktorý budem nazývať „falzifikovateľnosť“, je vedecky žiadúca vlastnosť hypotézy, aby sústredila svoju masu pravdepodobnosti do uprednostňovaných výsledkov, z čoho vyplýva, že zároveň musí priradiť nízku pravdepodobnosť nejakým neuprednostňovaným výsledkom; pravdepodobnosti musia dávať súčet 1, k dispozícii máme iba obmedzené množstvo pravdepodobnosti. V ideálnom prípade by mali existovať možné pozorovania, ktoré stláčajú pravdepodobnosť hypotézy takmer na nulu: Mali by existovať veci, ktoré táto hypotéza *nedokáže* vysvetliť, predstaviteľné experimentálne výsledky, s ktorými táto teória *nie je* zlučiteľná. Teória, ktorá dokáže vysvetliť všetko, nezakazuje nič, takže nám nedáva žiadnu radu, čo máme očakávať.

$$P(A_i|B) = P(B|A_i) \times P(A_i) / \text{Suma}_j\{ P(B|A_j) \times P(A_j) \}$$

→ http://lesswrong.com/lw/q3/decoherence_is_simple/

V pojmoch Bayesovej vety, ak existuje aspoň jedno také pozorovanie B, ktoré hypotéza Ai nedokáže vysvetliť, čiže P(B | Ai) je maličké, potom aj čitateľ P(B | Ai) × P(Ai) bude maličký, a teda aj výsledná pravdepodobnosť P(Ai | B). Aktualizácia na základe videnia nemožného výsledku B posunula pravdepodobnosť Ai nadol takmer na nulu. Teória, ktorá odmieta vystaviť sa takejto zraniteľnosti, bude musieť rozdeliť svoju pravdepodobnosť naširoko, aby nemala žiadne diery; nedokáže silno sústrediť pravdepodobnosť do niekoľkých uprednostňovaných výsledkov; nebude môcť poskytovať presné rady.

Toto je pravidlo vedy odvodené z teórie pravdepodobnosti.

Ako je tu znázornené, „falzifikovateľnosť“ je niečo, čo vyhodnocujete pohľadom na *jednu* hypotézu a otázkou: „Ako úzko sústreďuje svoju pravdepodobnostnú distribúciu pre možné výsledky? Ako úzko mi hovorí, čo mám očakávať? Dokáže vysvetliť, prečo sú niektoré výsledky omnoho lepšie než iné?“

Je interpretácia dekoherencie kvantovej fyzicky *falzifikovateľná*? Existujú experimentálne výsledky, ktoré by mohli znížiť jej pravdepodobnosť na infinitezimálnu?

Iste: Mohli by sme merať previazané častice, ktoré by vždy mali mať opačný spin, a zistiť, že ak ich odmeriame dostatočne ďaleko od seba, niekedy budú mať rovnaký spin.

Alebo by sme mohli zistiť, že jablká padajú nahor, že planéty slnečnej sústavy chodia po náhodných krivoľakých čiarach, alebo že nejaký atóm vyžaruje fotóny bez zrejmého zdroja energie. Aj tieto pozorovania by falzifikovali dekoherentnú kvantovú mechaniku. Sú to veci, ktoré by sme podľa hypotézy, že dekoherentná QM vládne vesmíru, jednoznačne *nemali očakávať*, že uvidíme.

Existujú teda pozorovania B, kde P(B | A_{deko}) je infinitezimálne, čo by znížilo P(A*deko* | B) na infinitezimálne.

> „Ale to je len preto, že dekoherentná kvantová mechanika je stále kvantová mechanika! Čo samotná tá časť o dekoherencii v porovnaní s predpokladom kolapsu?“

Dostávame sa tam. Pointa je, že som práve definoval test, ktorý vás vedie rozmýšľať vždy len o jednej hypotéze (a nazval som ho „falzifikovateľnosť“). Ak chcete odlíšiť dekoherenciu *od* kolapsu, musíte myslieť naraz aspoň na *dve* hypotézy.

V skutočnosti ten test „falzifikovateľnosti“ nie je až *tak* jednoznačne zameraný, veď súčet v menovateli musí obsahovať *nejaké* iné hypotézy. Ale to, čo som práve definoval ako „falzifikovateľnosť“, zachytáva ten typ problému, na ktorý sa Karl Popper sťažoval, keď povedal, že Freudovská psychoanalýza je „nefalzifikovateľná“, lebo bola rovnako dobrá vo vysvetľovaní ľubovoľnej veci, ktorú mohol pacient urobiť.

Keby ste patrili k mimozemskému druhu, ktorý nikdy nevymyslel predpoklad kolapsu ani copenhagenskú interpretáciu – keby jediná fyzikálna teória, o ktorej ste *kedy* počuli, bola dekoherentná QM – keby ste v hlave mali iba diferenciálne rovnice pre vývoj vlnovej funkcie plus Bornovo pravidlo pravdepodobnosti – stále by ste mali jasné očakávania o vesmíre. Nežili by ste v čarovnom svete, kde je možné všetko.

> „Ale to isté by si mohol povedať o kvantovej mechanike *bez* (makroskopickej) dekoherencie.“

Veď áno! Niekto, kto chodí okolo s diferenciálnou rovnicou pre vývoj vlnovej funkcie plus predpokladom kolapsu, ktorý sa riadi Bornovými pravdepodobnosťami a spúšťa sa predtým než superpozícia dosiahne makroskopické rozmery, tiež stále žije vo svete, kde jablká padajú nadol a nie nahor.

„Ale kde dáva dekoherencia *novú* predpoveď, ktorú by sme mohli *testovať*?“

Voči čomu „novú“ predpoveď? Voči stavu vedomostí, ktoré mali starovekí Gréci? Keby ste sa vrátili v čase a ukázali im dekoherentnú QM, umožnilo by im to robiť veľa experimentálnych predpovedí, ktoré predtým nevedeli urobiť.

Keď poviete „nová predpoveď“, myslíte tým „nová“ relatívne k nejakej inej hypotéze, ktorá definuje „starú predpoveď“. Toto nás dostáva k teórii toho, čo som sa rozhodol označiť

ako *testovateľnosť*; a tento algoritmus zo svojej podstaty naraz zvažuje aspoň dve hypotézy. Nemôžete niečo nazvať „*nová* predpoveď“, keď uvažujete izolovane iba o jednej odpovedi.

V bayesovských pojmoch, hľadáte kus indície B, ktorý poskytne indíciu v prospech jednej z hypotéz voči druhej, rozlíši medzi nimi, a tento proces vytvárania indície môžeme nazvať „test“. Pozeráte sa na experimentálny výsledok B, pre ktorý:

$$p(B \mid A_d) \mathrel{!}= p(B \mid A_c)$$

čiže nejaký výsledok B, ktorý má inú pravdepodobnosť podľa toho, či platí hypotéza dekoherencie alebo či platí hypotéza kolapsu. Z čoho potom vyplýva, že výsledné pravdepodobnosti pre dekoherenciu a kolaps sa budú odlišovať od pôvodných pravdepodobností.

$$p(B \mid A_d) / p(B \mid A_c) \mathrel{!}= 1, \text{ a preto}$$

$$p(Ad \mid B) / p(Ac \mid B) = p(B \mid Ad) / p(B \mid Ac) \times p(Ad) / p(Ac)$$

$$p(Ad \mid B) / p(Ac \mid B) \mathrel{!}= p(Ad) / p(Ac)$$

Táto rovnica je súmerná (za predpokladu, že sa žiadna pravdepodobnosť nerovná doslova 0). Nie je tu jedno Aj s názvom „stará hypotéza“ a iné Aj s názvom „nová hypotéza“.

Táto súmernosť je v teórii pravdepodobnosti zámer, nie chyba! Ak dizajnujete umelý rozmýšľací systém, ktorý dôjde k rôznym názorom podľa toho, v akom poradí mu predložia indície, to sa nazýva „hysterézia“ a považuje sa to za Zlú Vec. Počul som, že sa na to mračia aj vo Vede.

Z pohľadu teórie pravdepodobnosti máme rôzne triviálne vety, ktoré hovoria, že by nemalo záležať na tom, či najprv aktualizujete podľa X a potom podľa Y, alebo aktualizujete najprv podľa Y a potom podľa X. Prinajmenšom by boli triviálne, keby ich ľudia neporušovali tak často a tak bezstarostne.

Ak je dekoherencia „netestovateľná“ voči kolapsu, potom je rovnako kolaps „netestovateľný“ voči dekoherencii. Čo keby sa história fyziky odvíjala iným spôsobom – čo keby Hugh Everett a John Wheeler stáli na mieste Bohra a Heisenberga, a naopak? Bolo by vtedy správne a primerané, aby ľudia v danom svete pozreli na interpretáciu kolapsu, vzdychli a povedali: „Kde sú tu *nové* predpovede?“

Čo keby sme jedného dňa stretli mimozemský druh, ktorý vymyslel dekoherenciu pred kolapsom? Musí si každý z nás držať tú teóriu, ktorú vymyslel ako prvú? Nemá Rozum čo povedať k tejto téme, a nezanechá nám žiaden prostriedok, ako vyriešiť túto hádku, okrem medzihviezdnej vojny?

> „Ale ak sa vzdáme požiadavky na nové predpovede, zostane nám vedecký chaos. Môžete k starým teóriám pridávať ľubovoľné netestovateľné komplikácie, a dostanete experimentálne ekvivalentné predpovede. Ak odmietneme to, čo nazývaš ‚hysterézia‘, ako môžeme brániť naše terajšie teórie proti každému šarlatánovi, ktorý navrhne, že elektróny majú vlastnosť s názvom ‚vôňa‘ rovnako ako kvarky majú ‚chuť‘?“

Povedzme na úvod, že celkom súhlasím, že by ste mali odmietnuť toho, kto príde za vami a povie: „Hej, mám túto skvelú novú myšlienku! Možno nabité častice neťahá elektromagnetické pole. Možno existujú maličkí anjeli, ktorí v skutočnosti posúvajú tieto častice, a elektromagnetické pole im iba hovorí, ako to majú robiť. Pozrite, mám všetky tieto úspešné experimentálne predpovede – predpovede, ktoré ste kedysi považovali za svoje vlastné!“

Takže áno, súhlasím, že by sme nemali prijať túto úžasnú novú teóriu, ale jej problémom nie je to, že je *nová*.

Predstavte si, že by sa ľudská história odvíjala iba trochu odlišne, a že by Cirkev bola hlavnou grantovou agentúrou Vedy. A predstavte si, že keď sa prvýkrát vypracovali zákony elektromagnetizmu, považoval by sa jav magnetizmu za dôkaz existencie neviditeľných duchov, anjelov. Z Jamesa Clerka by sa stal Svätý Maxwell, ktorý opísal zákony riadiace konanie anjelov.

O pár storočí neskôr, keď už by sa moc Cirkvi upaľovať ľudí na hranici potlačila, niekto by prišiel a povedal: „Hej, naozaj potrebujeme tých anjelov?“

„Áno,“ povedali by všetci. „Ako inak by sa púhe čísla v elektromagnetickom poli premenili na skutočný pohyb častíc?“

„Možno je to základný zákon," povie nováčik, „alebo je to možno niečo iné ako anjeli, čo objavíme neskôr. Ja iba tvrdím, že interpretovať tie čísla *ako konanie anjelov* nám v skutočnosti nič nepridáva, a mali by sme si iba nechať tie čísla a vyhodiť tú časť s anjelmi."

A oni sa pozrú jeden na druhého a nakoniec povedia: „Ale tvoja teória nedáva žiadne nové experimentálne predpovede, prečo by sme si ju mali osvojiť? Ako otestujeme tvoje tvrdenia o neprítomnosti anjelov?"

Z normatívneho pohľadu sa mi zdá, že ak by sme mali odmietnuť anjelov šarlatána v tom prvom prípade, *hoci tieto dve teórie nedokážeme experimentálne odlíšiť*, mali by sme odmietnuť aj anjelov zavedenej vedy v tom druhom prípade, hoci tieto dve teórie nedokážeme experimentálne odlíšiť.

Zvyčajne sú to šarlatáni, kto pridáva nové zbytočné komplikácie, a nie vedci, ktorí by ich omylom zabudovali na začiatku. Ale problém nie je v tom, že tie komplikácie sú nové, ale že sú zbytočné bez ohľadu na to, či sú nové alebo nie.

Bayesovec by povedal, že tá komplikovanosť teórie s anjelmi navyše vedie k pokute za pôvodnú pravdepodobnosť teórie. Ak dve teórie robia ekvivalentné predpovede, ponecháme si tú, ktorú možno opísať pomocou najkratšej správy, najmenšieho programu. Ak vyhodnocujete pôvodnú pravdepodobnosť každej hypotézy počítaním bitov kódu a následným použitím bayesovskej aktualizácie podľa všetkých dostupných indícií, nie je žiaden rozdiel v tom, ktorú hypotézu ste počuli ako prvú, ani v poradí, v akom aplikujete indície.

V skutočnom živote zvyčajne nie je možné aplikovať formálnu teóriu pravdepodobnosti, rovnako ako nedokážete predpovedať víťaza tenisového zápasu pomocou kvantovej teórie poľa. Ale ak môže teória pravdepodobnosti poslúžiť praxi ako sprievodca, potom hovorí toto: „Odmietajte *zbytočné* komplikácie *vo všeobecnosti*, nie iba ak sú *nové*."

> „Áno, a *zbytočné* sú presne mnohé svety dekoherencie! Údajne všetky tieto svety existujú zároveň s naším, a nič v našom svete *nerobia*, ale ja v ne mám aj tak veriť?"

Nie, podľa dekoherencie to, čomu máte veriť, sú všeobecné zákony, ktorými sa riadia vlnové funkcie – a tieto všeobecné zákony sú veľmi viditeľné a testovateľné.

Na inom mieste som argumentoval, že <u>imprimatur vedy by sa malo spájať so všeobecnými zákonmi, nie s konkrétnymi udalosťami</u>, pretože tie všeobecné zákony sú to, čo v princípe hocikto môže ísť sám otestovať. Ubezpečujem vás, že ako píšem tieto slová, mám oblečené biele ponožky. Pravdepodobne teda máte *racionálne* zdôvodnenie, prečo veriť tomuto historickému faktu. Ale nie je to ten osobitne silný druh tvrdenia, aké kanonizujeme za dočasné názory vedy, pretože neexistuje žiaden experiment, ktorý by ste mohli sami urobiť, aby ste zistili jeho pravdu; ste odkázaní na moju autoritu. Keby som vám ale povedal o hmotnosti elektrónov všeobecne, mohli by ste si ísť pohľadať vlastný elektrón na testovanie, a tak by ste sami videli pravdivosť tohto všeobecného zákona na tom konkrétnom príklade.

Táto schopnosť hocikoho ísť a sám si overiť všeobecný vedecký zákon tým, že si vytvorí konkrétny príklad, je to, čo robí našu vieru vo všeobecný zákon zvlášť spoľahlivou.

Dekoherentisti hovoria, že veria v diferenciálnu rovnicu, ktorá podľa pozorovaní riadi vývoj vlnových funkcií – môžete ísť a otestovať si to sami kedykoľvek sa vám zachce; skrátka sa pozrite na atóm vodíka.

Viera v existenciu oddelených úsekov všeobecnej vlnovej funkcie nie je *dodatočná*, a nemá *za úlohu* vysvetliť cenu zlata v Londýne; je to iba deduktívny dôsledok vývoja vlnovej funkcie. Ak vám indície mnohých konkrétnych príkladov dávajú dôvod veriť, že X -> Y je všeobecný zákon, a ak vám indície niektorého konkrétneho príkladu dávajú dôvod veriť v X, potom by ste mali mať $P(Y) \geq P(X \text{ a } (X \rightarrow Y))$.

Alebo z iného pohľadu, ak $P(Y \mid X) \approx 1$, potom $P(X \text{ a } Y) \approx P(X)$.

Čo znamená, že veriť detailom navyše vás nestojí pravdepodobnosť navyše, ak sú *logickými dôsledkami* všeobecných názorov, ktoré už máte. Predpokladáme však, že samotné tie všeobecné názory sú falzifikovateľné, lebo inak načo sa s nimi unúvať?

To je dôvod, prečo neveríme, že vesmírna loď zmizne z existencie, keď prekročí kozmologický horizont relatívne voči nám. Je pravda, že ďalšia existencia tejto vesmírnej lode nemá vplyv na náš svet. Ďalšia existencia tejto vesmírnej ode nám nepomáha vysvetliť ceny zlata v Londýne. Ale túto vesmírnu loď dostaneme zadarmo ako dôsledok všeobecných zákonov, ktoré naznačujú zachovanie hmoty a energie. Keby ďalšia existencia vesmírnej lode *nebola* deduktívnym dôsledkom zákonov fyziky, ako ich dnes modelujeme, *potom* by to bola dodatočná podrobnosť, stála by pravdepodobnosť navyše, a my by sme sa museli pýtať, prečo naša teória musí obsahovať toto tvrdenie.

Tá časť dekoherencie, ktorá má byť testovateľná, nie sú mnohé svety ako také, ale iba všeobecný zákon, ktorým sa riadi vlnová funkcia. Dekoherentisti poznamenávajú, že ak tento zákon platí všeobecne, vyplýva z neho existencia celých svetov v superpozícii. Existuje kritika, ktorú možno proti tejto teórii namieriť, najznámejšie je: „Ale z čoho vychádzajú Bornove pravdepodobnosti?“ Avšak v rámci vnútornej logiky dekoherencie sa mnohé svety neponúkajú ako vysvetlenie ničoho, ani nie sú podstatou teórie, ktorú treba otestovať; sú jednoducho logickým dôsledkom tých všeobecných zákonov, ktoré tvoria podstatu tejto teórie.

Ak A -> B, potom ~B -> ~A. Aby sme popreli existenciu svetov v superpozícii, museli by sme poprieť univerzálnosť kvantových zákonov formulovaných v súlade s fungovaním atómov vodíka a všetkých ďalších preskúmateľných prípadov; a práve toto popieranie pripadá dekoherentistom ako ten netestovateľný detail navyše. Zvyšné časti vlnovej funkcie nemôžete vidieť – načo mať *navyše* predpoklad, že neexistujú?

Udalosti obklopujúce kontroverziu dekoherencie sú možno jedinečné v histórii vedy, označujú prvý moment, keď vážni vedci prišli a povedali, že historickou zhodou okolností ľudstvo vyvinulo mocnú, úspešnú, matematickú teóriu fyziky, ktoré obsahuje anjelov. Že existuje celý zákon, predpoklad kolapsu, ktorý možno jednoducho odhodiť, čím teória zostane *striktne* jednoduchšia.

Do tejto diskusie by som rád prispel tvrdením, že vo svetle matematicky solídneho chápania teórie pravdepodobnosti, dekoherencia nie je vylúčená Occamovou britvou, a nie je ani nefalzifikovateľná, ani netestovateľná.

Môžeme zvážiť napríklad dekoherenciu a predpoklad kolapsu vedľa seba, a vyhodnocovať kritiky ako: „Nepredpokladá dekoherencia, že kvantové pravdepodobnosti by mali byť vždy 50/50?“ a „Neporušuje kolaps špeciálnu relativitu tým, že predpokladá pôsobenie na diaľku?“ Môžeme zvažovať relatívne zásluhy týchto teórií na základe ich kompatibility s vnemami a so zdanlivou povahou fyzikálneho zákona.

Tvrdiť, že dekoherencia nie je ani len pripustená do hry – pretože samotné mnohé svety sú „veci navyše“, ktoré porušujú Occamovu britvu, alebo pretože mnohé svety sú „netestovateľné“, alebo pretože dekoherencia nerobí žiadne „nové predpovede“ - toto všetko je, podľa mňa, jasná chyba z pohľadu teórie pravdepodobnosti. V diskusii by sme mali tieto konkrétne argumenty jednoducho škrtnúť a ísť ďalej.

*→

236. Uprednostňovanie hypotézy

Predstavte si, že polícia v Largeville, meste s miliónom obyvateľov, vyšetruje vraždu, v ktorej sú len malé náznaky, alebo žiadne – obeť bola dobodaná na smrť v uličke, nie sú žiadne odtlačky ani svedkovia.

Potom povie jeden z detektívov: „Nuž... nemáme žiadnu predstavu, kto to urobil... žiadnu konkrétnu indíciu vyberajúcu niekoho z milióna ľudí v tomto meste... ale *zamyslime sa nad hypotézou*, že túto vraždu spáchal Mortimer Q. Snodgrass žijúci na Obyčajnej ulici číslo 128. Napokon, *mohol* to byť aj on.“

Toto by som označil ako *klam uprednostňovania hypotézy*. (Povedzte mi, ak to už má nejaké oficiálne meno – nespomínam si, že by som to videl opísané.)

→ http://lesswrong.com/lw/q4/decoherence_is_falsifiable_and_testable/

Možno tento detektív má nejaký druh rozumnej indície, ktorá nie je legálna indícia prípustná na súde – napríklad klebetu od informátora. Ale ak tento detektív nemá *už poruke žiadne* zdôvodnenie, prečo vyzdvihuje Mortimera do osobitnej pozornosti polície – ak toto meno celkom náhodne vytiahol z klobúka – potom boli porušené Mortimerove práva.

A toto platí aj keď daný detektív netvrdí, že to Mortimer „urobil“, akurát žiada políciu, aby trávila čas uvažovaním, že to Mortimer *mohol* urobiť – bezdôvodne vyzdvihuje do pozornosti túto konkrétnu hypotézu. Je v ľudskej povahe hľadať potvrdenie namiesto vyvrátenia. Predstavme si, že traja detektívi navrhnú každý svojho nenávideného nepriateľa ako mená, ktoré treba zvážiť; a že Mortimer má hnedé vlasy, Frederick má čierne vlasy, a Helen je blondína. Potom sa nájde svedok, ktorý povie, že osoba odchádzajúca z miesta činu mala hnedé vlasy. „Aha!“ povie polícia. „Predtým sme nemali žiadne indície ako rozlíšiť medzi možnosťami, ale *teraz* vieme, že to urobil Mortimer!“

Toto súvisí s princípom, ktorý som začal nazývať „vyhľadanie hypotézy“, ktorý znie, že ak máte miliardu krabíc, z ktorých iba jedna obsahuje diamant (pravdu), a vaše detektory dávajú každý iba 1 bit indície, potom treba omnoho viac indície na to, aby sa pravda vyzdvihla do vašej osobitnej pozornosti – aby sa to zúžilo na desať dobrých možností, z ktorých každá si zaslúži, aby sme si ju všimli jednotlivo – než na zistenie, *ktorá* z týchto desiatich možností je pravdivá. 27 bitov, aby sme to zúžili na 10, a potom iba 4 ďalšie bity nám dajú prevažnú šancu na správnu odpoveď.

Preto ten detektív tým, že bezdôvodne vyzdvihol Mortimera do osobitnej pozornosti polície spomedzi milióna iných ľudí, preskočil *väčšinu indície*, ktorú proti Mortimerovi treba dodať.

A ten detektív by mal mať túto indíciu u seba v prvej chvíli, keď na Mortimera *vôbec* upozorní políciu. Môže to byť iba racionálna indícia namiesto legálnej indície, ale ak neexistuje *žiadna indícia*, potom tento detektív chudáka Mortimera obťažuje a prenasleduje.

Počas môjho nedávneho diavlogu so Scottom Aaronsonom o kvantovej mechanike sa mi podarilo Scotta zatlačiť do kúta natoľko, že uznal, že neexistuje absolútne žiadna konkrétna indícia, ktorá by uprednostňovala predpoklad kolapsu alebo kvantovú mechaniku jedného sveta. Ale, povedal Scott, možno sa s indíciou v prospech kvantovej mechaniky jedného sveta stretneme *v budúcnosti*, a mnohé svety majú stále otvorenú otázku Bornových pravdepodobností.

Tak toto je to, čo by som označil za klam uprednostňovania hypotézy. Musí existovať bilión lepších spôsobov, ako odpovedať na Bornovu otázku bez pridávania predpokladu kolapsu, ktorý by bol jediným nelineárnym, neunitárnym, nespojitým, nediferencovateľným, CPT-nesymetrickým, nelokálnym v priestore konfigurácií, porušujúcim Liovilleovu vetu, obsahujúcim privilegovaný priestor súčasnosti, ovplyvňujúcim rýchlejšie než svetlo, nekauzálnym, neformálne definovaným zákonom v celej fyzike. Niečo takéto nefyzikálne si nezaslúži, aby sa o tom *nahlas hovorilo*, by ani len *uvažovalo ako o možnosti*, pokiaľ to nemá extra veľkú hmotnosť indícií – omnoho väčšiu než je dnešný celkový počet nula.

Ale vďaka historickej nehode má predpoklad kolapsu a kvantovú mechaniku jedného sveta predsa len každý v ústach a na mysli, a tak sa otvorená otázka Bornových pravdepodobností ponúka (nikým menším než je Scott Aaronson!) ako indícia, že mnohé svety zatiaľ nevedia ponúknuť úplný obraz sveta. Čo má akože znamenať, že QM jedného sveta je stále akosi v hre.

V mysliach ľudských bytostí, ak ich dokážete primäť, aby mysleli na túto konkrétnu hypotézu namiesto biolóna iných možností, ktoré nie sú o nič zložitejšie ani nepravdepodobnejšie, ste naozaj *urobili* obrovský kus práce v presviedčaní. Čokoľvek, o čom rozmýšľajú, vnímajú ako „v hre“, a ak iní hráči napohľad v preteku o čosi zaostanú, predpokladajú, že tento hráč sa posúva vpred, či nebodaj do čela.

A áno, presne toho istého klamu sa dopustí, v omnoho zrejmejšej mierke, veriaci, ktorý podotkne, že moderná veda neponúka absolútne úplné vysvetlenie celého vesmíru, a berie toto ako indíciu, že existuje Jehova. A nie Alah, lietajúce špagetové monštrum, alebo bilión iných o nič zložitejších bohov – nehovoriac o celom priestore naturalistických vysvetlení!

Hovoriť o „inteligentnom dizajne“ kedykoľvek ukážete na údajnú chybu alebo otvorený problém v evolučnej teórii, je opäť uprednostňovaním hypotézy – musíte mať *už v ruke* indíciu, ktorá ukazuje

konkrétne na inteligentný dizajn, aby ste zdôvodnili, prečo *do našej pozornosti vyzdvihujete túto konkrétnu myšlienku* namiesto tisíca iných.

Takže toto je to *príčetné* pravidlo. A zodpovedajúca anti-epistemológia je donekonečna hovoriť o „možnosti“ a o tom, ako „nemôžete vyvrátiť“ nejakú myšlienku, dúfať, že budúca indícia to môže potvrdiť, keď v súčasnosti nemáte žiadnu indíciu v ruke, nástojiť a nástojiť na *možnostiach* bez vyhodnocovania možnej nesúhlasiacej indície, kresliť žiarivé slovné obrázky potvrdzujúcich pozorovaní, ktoré by sa *mohli* stať, ale *ešte* sa nestali, alebo pokúšať sa ukázať, že kúsok za kúskom negatívnej indície „nie je definitívny“.

Tak ako Occamova britva hovorí, že zložitejšie výroky vyžadujú viac indície, aby sme im verili, zložitejšie výroky by si mali vyžadovať viac práce, aby sme im venovali pozornosť. Rovnako ako princíp priťažujúcich podrobností vyžaduje, aby každá časť názoru bola samostatne zdôvodnená, vyžaduje si, aby každá časť bola samostatne vyzdvihnutá do pozornosti.

Ako sme si povedali v Názoroch ako večný pohyb, viera a stroje na večný pohyb 2. typu (voda -> kocky ľadu + elektrina) majú spoločné to, že budia dojem, že *vyrábajú nepravdepodobnosť z ničoho*, či už je to nepravdepodobnosť vody vytvárajúcej kocky ľadu alebo nepravdepodobnosť dôjdenia k správnym názorom bez pozorovania. Niekedy väčšina anti-práce vo výrobe tejto nepravdepodobnosti spočíva v tom, že nás niekto privedie, aby sme *venovali pozornosť* neodôvodnenému názoru – myslenie naň, nástojenie na ňom. V širokom priestore odpovedí je pozornosť bez indície viac než polovica cesty k názoru bez indície.

Niekto, kto trávi celý deň rozmýšľaním, či *svätá Trojica* existuje alebo neexistuje, namiesto či existuje Alah alebo Thor alebo Lietajúce Špagetové Monštrum, je viac než na polceste ku kresťanstvu. Ak odchádza, odišiel menej než napoly; ak prichádza, už je tam viac než z polovice.

Často sa vyskytujúci spôsob uprednostňovania je pokúšať sa neistotu vnútri nejakého priestoru preliať mimo tohto priestoru do uprednostňovanej hypotézy. Napríklad kreacionista sa chytí nejakej (údajne) diskutovanej stránke súčasnej teórie, povie, že vedci si *nie sú istí ohľadom evolúcie*, a potom povie: „V skutočnosti nevieme, ktorá teória je správna, takže možno je správny inteligentný dizajn.“ Lenže daná neistota bola neistotou *v rámci* oblasti prirodzených teórií evolúcie – nemáme žiaden dôvod veriť, že potrebujeme opustiť toto územie, aby sme si s našou neistotou poradili, a ešte *menší* dôvod vyskočiť z oblasti štandardnej vedy a dopadnúť *konkrétne na Jehovu*. To je uprednostňovanie hypotézy – branie pochybnosti *v* normálnom priestore a pokúšanie sa preliať túto pochybnosť z normálneho priestoru do uprednostňovaného (a zvyčajne diskreditovaného) *extrémne* nenormálneho cieľa.

Podobne, naša neistota o tom, odkiaľ pochádzajú Bornove štatistiky, by mala byť neistotou *v rámci* priestoru kvantových teórií, ktoré sú spojité, lineárne, unitárne, pomalšie než svetlo, lokálne, kauzálne, naturalistické, a tak ďalej – zvyčajná povaha fyzikálneho zákona. Niečo z tejto neistoty by sa mohlo vyliať mimo štandardnho priestoru do teórií, ktoré porušujú *jednu* z týchto štandardných vlastností. Je naozaj možné, že by sme mali uvažovať mimo vychodených koľají. Ale teórie jedného sveta porušujú *všetky* tieto vlastnosti, a neexistuje dôvod uprednostňovať túto hypotézu.

*→

237. Život v mnohých svetoch

Niektorí diskutéri nedávno vyjadrili znepokojenie z predstavy, ako sa neustále rozdeľujú na zilióny ľudí, čo je priamočiara a nevyhnutná predpoveď kvantovej mechaniky.

Iní priznali, že si nie sú istí, aké sú dôsledky mnohých svetov na plánovanie: Ak sa rozhodujete zapnúť si v tomto svete bezpečnostný pás, zvyšuje to šancu, že vaše iné ja si svoj bezpečnostný pás nezapne? Ste sebecký na ich úkor?

Pamätajte na Eganov zákon: *To všetko dokopy tvorí normálnosť.*

→ http://lesswrong.com/lw/19m/privileging_the_hypothesis/

(Autor: Greg Egan, *Karanténa*.[208])

Frank Sulloway povedal:[209]

> Ironicky, výhodou psychoanalýzy oproti Darwinizmu je práve to, že jej predpovede sú také bizarné a jej vysvetlenia sú také antiintuitívne, že si pomyslíme: *Je to naozaj tak? Aké radikálne!* Freudove myšlienky sú také zaujímavé, že sú za ne ľudia ochotní platiť, zatiaľ čo jedna z veľkých nevýhod Darwinizmu je, že máme pocit, že to už vieme, pretože v istom zmysle to naozaj už vieme.

Keď Einstein zvrhol Newtonovskú verziu gravitácie, jablká neprestali padať, planéty neodbočili do Slnka. Každá nová fyzikálna teória musí *zachytiť* úspešné predpovede starej teórie, ktorú nahradila; mala by predpovedať, že obloha bude modrá, nie zelená.

Nemyslite si teda, že mnohé svety sú to na to, aby robili zvláštne, radikálne, vzrušujúce predpovede. Všetko dokopy to tvorí normálnosť.

Prečo by to teda malo niekoho zaujímať?

Pretože sme si raz položili túto otázku, pre racionalistu fascinujúcu: *Čo* všetko tvorí dokopy normálnosť?

A ukázalo sa, že odpoveď na túto otázku znie: kvantová mechanika. Je to *kvantová mechanika*, čo dokopy tvorí normálnosť.

Keby existovalo niečo iné *namiesto* kvantovej mechaniky, *potom* by svet vyzeral čudne a nezvyčajne.

Pamätajte na toto, keď premýšľate, ako žiť v tom čudnom novom vesmíre s mnohými svetmi: *Vždy ste tam boli.*

Antropológovia nám hovoria, že náboženstvá zvyčajne vykazujú vlastnosť, ktorá sa volá *minimálna antiintuitívnosť*; sú dosť prekvapujúce na to, aby sme si ich zapamätali, ale nie také čudesné, aby sa *ťažko* pamätali. Anubis má hlavu ako pes, vďaka čomu si ho zapamätáme, ale zvyšok tela má ako človek. Duchovia dokážu vidieť cez steny; ale stále dostávajú hlad.

Lenže fyzika nie je náboženstvo nastavené, aby vás prekvapilo presne tak, aby ste si ho zapamätali. Javy v pozadí sú také antiintuitívne, že ľuďom trvá dlhé roky štúdia, než ich pochopia. Lenže javy na povrchu sú celkom bežné. *Nikdy* nezachytíte záblesk iného sveta kútikom oka. Nikdy nezačujete hlas nejakého svojho iného ja. Toto zákony jednoznačne vyslovene zakazujú. Smola, ste iba schizofrenik.

Akt *rozhodovania sa* nemá žiadnu špeciálnu interakciu s procesom, ktorý rozvetvuje svety. Vo vašej *mysli*, vo vašej *predstave*, vyzerá rozhodnutie ako bod rozvetvenia, z ktorého by svet mohol ísť dvoma cestami. Lenže by ste cítili rovnakú neistotu, predstavovali si tie isté alternatívy, aj keby existoval iba jeden svet. To je to, čo si ľudia mysleli staročia pred kvantovou mechanikou, a aj tak si predstavovali alternatívne výsledky, ktoré by mohli nastať z ich rozhodnutí.

Rozhodovanie a *dekoherencia* sú *celkom ortogonálne pojmy*. Keby sa váš mozog nikdy nestal dekoherentný, potom by si tento jediný kognitívny proces stále musel predstavovať rôzne voľby a ich rôzne dôsledky. Kameň, ktorý nerobí žiadne rozhodnutia, sa riadi tými istými zákonmi kvantovej mechaniky ako všetko ostatné, a horúčkovito sa rozdvojuje zatiaľ čo leží na jednom mieste.

Vy sa nerozdeľujete *vtedy, keď prídete ku konkrétnemu rozhodnutiu*, rovnako ako sa nerozdeľujete vtedy, keď sa nadýchnete. Vy sa skrátka rozdeľujete po celý čas v dôsledku dekoherencie, čo s rozhodovaním nemá nič spoločné.

Existuje populácia svetov, a v každom svete do všetko dokopy dáva normálnosť: jablká neprestanú padať. V každom svete si ľudia vyberajú tú cestu, ktorá im pripadá najlepšia. Možno im napadne iná myšlienková línia, a uvidia nové dôsledky, alebo im uniknú iné, a dôjdu k rozdielnemu záveru. Ale nie je to tak, že si v každom svete vyberiete inú možnosť. Nie je to tak, že si jedna verzia vás vyberie to, čo vyzerá najlepšie, a iná verzia si vyberie to, čo vyzerá najhoršie. V každom svete jablká ďalej padajú, a ľudia si ďalej vyberajú to, čo vyzerá ako dobrý nápad.

---

208 Greg Egan, *Quarantine* (London: Legend Press, 1992).

209 Robert S. Boynton, „The Birth of an Idea: A Profile of Frank Sulloway,“ *The New Yorker* (October 1999).

Áno, môžete vyhľadávať výnimky z tohto pravidla, ale sú to *normálne* výnimky. Toto všetko dokopy tvorí normálnosť, v každom svete.

Nemôžete si „vybrať, v ktorom svete skončíte“. Vo všetkých svetoch voľby ľudí určujú výsledky rovnakým spôsobom, ako keby existoval iba jediný svet.

Výber, ktorý urobíte tu, nemá žiaden zvláštny vyvažujúci účinok na nejaký svet inde. Neexistuje kauzálna komunikácia medzi dekoherovanými svetmi. V každom svete voľby ľudí ovplyvňujú budúcnosť toho sveta, nie nejakého iného sveta.

Ak si viete predstaviť rozhodovanie v jednom svete, viete si predstaviť rozhodovanie v mnohých svetoch: skrátka sa ten svet ustavične rozdvojuje, zatiaľ čo inak sa riadi presne tými istými pravidlami.

V žiadnom svete sa dva plus dva nerovná piatim. V žiadnom svete nemôže vesmírna loď letieť rýchlejšie než svetlo. Všetky kvantové svety sa riadia našimi zákonmi fyziky; ich existenciu v prvom rade potvrdzujú naše zákony fyziky. Od samého začiatku sa nestala jediná nezvyčajná vec, ani v tomto, ani v žiadnom inom svete. Všetky sa riadia zákonmi.

Existujú príšerné svety, ktoré sú naprosto mimo vašej schopnosti ovplyvňovať? Iste. Aj počas 12. storočia sa stali príšerné veci, ktoré sú mimo vašej schopnosti ovplyvňovať. Lenže 12. storočie nie je vaša zodpovednosť, lebo sa to, ako sa hovorí, „už stalo“. Odporúčam, aby ste považovali každý svet, ktorý nie je vo vašej budúcnosti, za časť „zovšeobecnenej minulosti“.

Žite vo svojom svete. Predtým ako ste vedeli o kvantovej fyzike, by vás nepokúšalo skúšať žiť vo svete, ktorý napohľad neexistuje. Vaše rozhodnutia by mali dokopy dávať tú istú normálnosť: nemali by ste skúšať žiť v kvantovom svete, s ktorým nemôžete komunikovať.

Vaše rozhodovacia teória by mala byť (takmer vždy) rovnaká, či predpokladáte, že je šanca 90 %, že sa niečo stane, alebo že sa to stane v 9 z 10 svetov. Ale, pretože ľudia majú ťažkosti pri predstavovaní si pravdepodobností, môže byť užitočné predstaviť si, že sa niečo stane v 9 z 10 svetov. Ale toto vám iba pomáha používať normálnu teóriu rozhodovania.

Ak ste to predtým odkladali, teraz je ten čas naučiť sa <u>držať hubu a počítať</u>. Ako som napísal v kapitole <u>Lotérie: Plytvanie nádejou</u>:

> Ľudský mozog nerobí 64-bitovú aritmetiku s plávajúcou desatinnou čiarkou, a nedokáže oslabiť emocionálnu silu príjemného očakávania o činiteľ 0,000 000 01 bez úplného zahodenia danej myšlienkovej línie.

A v <u>Novej vylepšenej lotérii</u>:

> Rozdiel medzi nulovou šancou na zbohatnutie a epsilonovou šancou je rádovo epsilon.
>
> Ak o tom pochybujete, nech sa epsilon rovná $1/10^{10^{100}}$.

Ak mysíte na svet, ktorý by mohol vzniknúť zákonitou cestou, ale ktorého pravdepodobnosť je trilión k jednej, a v tomto svete sa deje niečo veľmi príjemné alebo veľmi hrozné... nuž, ak je zákonitý, tak pravdepodobne existuje. Ale mali by ste sa pokúsiť vo svojich centrách odmeny alebo centrách averzie uvoľniť triliónkrát menej neurotransmiterov, aby ste tento svet *primerane* zvážili pri svojich rozhodnutiach. Ak si myslíte, že to nedokážete... neunúvajte sa naň myslieť.

Inak by ste rovnako dobre mohli ísť a kúpiť si žreb lotérie používajúcej kvantovo vygenerované náhodného číslo, čo je stratégia, ktorá má *zaručený* výsledok veľmi drobného mega-víťazstva.

Alebo je tu ďalší spôsob, ako na to myslieť: Zvažujete vynaloženie nejakej myšlienkovej energie na svet, ktorého frekvencia vo vašej budúcnosti je menej ako jedna bilióntina? Potom si kúpte 10-stennú kocku vo svojom miestnom obchode s hrami, a skôr než začnete myslieť na tento zvláštny svet, začnite hádzať kockou. Ak na kocke padne 9 dvanásťkrát po sebe, *potom* môžete rozmýšľať o tomto svete. Inak nestrácajte čas; čas na myslenie je obmedzený zdroj, ktorý treba rozumne využívať.

Môžete hodiť kockami koľkokrát chcete, ale nesmiete na daný svet myslieť, dokiaľ vám nepadne 9 dvanásťkrát po sebe. Potom naň môžete asi minútu myslieť. Potom musíte zase začať hádzať kockami.

Toto vám môže pomôcť uchopiť pojem „bilión k jednej“ na intuitívnejšej úrovni.

Ak sa v nejakom bode pristihnete pri myšlienke, že kvantová fyzika môže mať nejaký zvláštny, *nenormálny* dôsledok na každodenný život – tak by ste pravdepodobne mali presne tam prestať.

Ach, existuje *pár* dôsledkov mnohých svetov na etiku. Priemerné utilitariánstvo zrazu vyzerá omnoho príťažlivejšie – nemusíte sa starať o vytvorenie toľkých ľudí, koľko len môžete, pretože priestor ľudí už preskúmava celý kopec ľudí. Vy iba chcete, aby priemerná kvalita života bola čo najvyššia, v budúcich svetoch, ktoré sú vašou zodpovednosťou.

A vždy by ste mali mať radosť z objavu, pokiaľ ste *vy osobne* niečo nevedeli. Je nezmyselné hovoriť o tom, kto je „prvý“ alebo „jediný“ človek, ktorý niečo vie, keď všetko poznateľné je známe vo svetoch, ktoré nie sú ani vo vašej minulosti, ani v budúcnosti, a nie sú ani pred vami ani za vami.

Ale vo všeobecnosti to celé dáva dokopy normálnosť. Ak je vaše chápanie mnohých svetov čo len trochu *neisté*, a zvažujete, či veriť nejakému zvláštnemu návrhu, alebo cítite nejakú zvláštnu emóciu, alebo plánujete nejakú zvláštnu stratégiu, potom vám môžem dať veľmi jednoduchú radu: Nerobte to.

Kvantový vesmír nie je nejaké zvláštne miesto, do ktorého ste boli vrhnutí. Je to spôsob, ako veci vždy boli.

*→

---

[208]
[209]

238. Kvantový nerealizmus

„Existuje mesiac, keď sa naň nikto nepozerá?“

– Albert Einstein, pýtajúci sa Nielsa Bohra

Predstavte si, že by ste práve začali pracovať na teórii kvantovej mechaniky.

Začali by ste robiť pokusy, ktoré dávajú rôzne výsledky podľa toho, ako podrobne ich pozorujete. Začnete kopať pod známym povrchom skutočnosti a nájdete extrémne presný matematický popis, ktorý vám dá iba relatívnu frekvenciu výsledkov; a čo je horšie, skladá sa z komplexných čísel. Veci sa správajú v pondelok ako častice, a v utorok ako vlny.

Správnu odpoveď nemáte ani vo forme hypotézy, pretože nebude vymyslená ešte ďalších tridsať rokov.

Čo je to najlepšie, čo by ste *mohli* urobiť v takejto kaši?

Najlepšie, čo môžete urobiť, je *strohá* interpretácia kvantovej mechaniky: „drž hubu a počítaj“. Ďalej sa budete *snažiť* vyvíjať nové teórie, pretože robiť, čo môžete, neznamená vzdať sa. Ale predpokladáme, že správna odpoveď nebude k dispozícii tridsať rokov, čo znamená, že žiadna z týchto nových teórií nebude na nič dobrá. Robiť to *najlepšie*, čo môžete, by znamenalo uvedomiť si to, napriek tomu, že by ste hľadali cesty, ako otestovať tieto hypotézy.

To najlepšie, čo by ste teoreticky mohli urobiť, by *nezahŕňalo* hovorenie vecí ako: „Vlnová funkcia nám dáva iba pravdepodobnosti, nie istoty.“ Keď sa obzrieme, toto bolo skákanie k záveru; vlnová funkcia nám dáva istotu existencie mnohých svetov. Takže tá časť, že vlnová funkcia je iba pravdepodobnosť, nebola celkom správna. Počítali ste, ale zabudli ste držať hubu.

Keby ste robili to *najlepšie*, čo sa dalo bez dostupnej správnej odpovede, potom by sa po počutí o dekoherencii ukázalo, že ste nepovedali *nič* nezlučiteľné s dekoherenciou. Dekoherenciu nevylučujú údaje ani výpočty. Ak teda odmiernete potvrdiť ako pozitívne poznanie ľubovoľný výrok, ktorý si nevyžadujú údaje ani výpočty, výpočty vás nebudú *nútiť* povedať nič nezlučiteľné s dekoherenciou. Ani s čímkoľvek, čo by mohla byť správna teória, keby to nebola dekoherencia. Ak zídete z cesty, musí to byť z vášho vlastného podnetu.

Ale pre ľudí je ťažké držať hubu a počítať – *naozaj* držať hubu a počítať. Máme ohromný sklon považovať svoju nevedomosť za pozitívne poznanie.

→ http://lesswrong.com/lw/qz/living_in_many_worlds/

Neviem, či sa niekedy naozaj odohral takýto rozhovor, ale takto sa z nevedomosti stáva poznanie:

Salviati: „Drž hubu a počítaj.“

Simplicio: „Prečo?“

Salviati: „Pretože nevieme, čo tieto rovnice znamenajú, akurát sa zdá, že fungujú.“

...o päť minút neskôr...

Simplicio: „Drž hubu a počítaj.“

Študent: „Prečo?“

Simplicio: „Pretože tieto rovnice nič *neznamenajú*, akurát fungujú.“

Študent: „Naozaj? Ako to vieš?“

Simplicio: „Povedal mi to Salviati.“

Podobná premena sa stane pri skoku z:

Salviati: „Keď moje výpočty ukazujú aplitúdu -1/3i pre pohltenie tohto fotónu, moje pokusy ukázali, že fotón bol pohltený okolo 107-krát z 1000, čo sa celkom podobá na 1/9, druhú mocninu absolútnej hodnoty. Je jasné, že existuje nejaká súvislosť medzi experimentálnymi štatistikami a druhou mocninou absolútnej hodnoty, ale neviem, aká.“

Simplicio: „Amplitúda pravdepodobnosti nehovorí, kde ten elektrón *je*, ale kde by *mohol byť*. Druhá mocnina absolútnej hodnoty je pravdepodobnosť, že sa skutočnosť ukáže takto. Skutočnosť *samotná* je nedeterministická.“

A zase:

Salviati: „Keď raz niečo odmeriam a dostanem experimentálny výsledok, svoje budúce výpočty robím iba s pomocou amplitúdy, ktorej druhá mocnina absolútnej hodnoty išla do výpočtu frekvencie tohto experimentálneho výsledku. Iba pri tomto pravidle moje nasledujúce výpočty zodpovedajú pozorovaným frekvenciám.“

Simplicio: „Keďže amplitúda *je* pravdepodobnosť, akonáhle *poznáte* experimentálny výsledok, pravdepodobnosť všetkého ostatného sa zmení na nulu!“

Celé toto skĺznutie z:

Druhá mocnina tejto „amplitúdy“ tesne súvisí s našimi experimentálne pozorovanými frekvenciami

cez

Amplitúda je pravdepodobnosť, že niečo nameriame

do

No samozrejme, že keď sme niečo nenamerali, pravdepodobnosť sa zmení na nulu

musí byť jedna z najzahanbujúcejších nesprávnych odbočiek v dejinách vedy.

Ak toto všetko vezmete *doslovne*, stane sa z toho vedomie-spôsobuje-kolaps interpretácia kvantovej mechaniky. V dnešnej dobe sa už asi nikto neprizná, že by *naozaj* veril v to, že vedomie spôsobuje kolaps kvantovej mechaniky...

Ale učebnice fyziky to stále takto píšu! Ľudia hovoria, že tomu neveria, ale hovoria, *ako keby* vedomosti boli zodpovedné za odstránenie nezlučiteľných amplitúd „pravdepodobnosti“.

Napriek tomu, akokoľvek nepravdepodobné mi pripadá, že vedomie spôsobuje kolaps, dáva nám to aspoň obrázok skutočnosti. Iste, je to neformálny obrázok. Iste, dáva myšlienkovým vlastnostiam ontologicky základné postavenie. Nemôžete *vypočítať*, kedy nastane „pozorovanie experimentu“ alebo kedy ľudia „vedia“, vy *skrátka viete*, kedy sú isté pravdepodobnosti *samozrejme* nulové. A toto „skrátka viete“ zhodou okolností zodpovedá vašim experimentálnym výsledkom, nech sú akékoľvek...

...ale prinajmenšom vedomie-spôsobuje-kolaps budí dojem, že nám hovorí, ako funguje vesmír. Amplitúdy sú skutočné, kolaps je skutočný, vedomie je skutočné.

Porovnajte si to s touto schémou argumentácie:

Študent: „Počkaj, hovoríš, že táto amplitúda zmizne, akonáhle mi meranie povie, že to nie je pravda?“

Simplicio: „Nie, nie! Nič tam *doslova* nemizne. Tie rovnice nič neznamenajú – oni iba dávajú dobré predpovede.“

Študent: „Ale čo sa potom *deje*?“

Simplicio: *(Sipí. Syčí.)* „Na toto sa nikdy nepýtaj.“

Študent: „A čo tá časť, kde tu odmeriame polarizáciu fotónu, a o svetelný rok ďalej sa pravdepodobnosť zvislej polarizácie previazaného fotónu zmení z 50 % na 25 %?“

Simplicio: „Áno, čo s tým?“

Študent: „To neporušuje špeciálnu relativitu?“

Simplicio: „Nie, pretože si sa iba *dozvedel* o polarizácii toho druhého fotónu. Pamätaj si, že amplitúdy nie sú *skutočné*.“

Študent: „Ale Bellova veta ukazuje, že nemôže byť žiadna lokálna skrytá premenná, ktorá by opisovala polarizáciu fotónu predtým, než ju nameriame...“

Simplicio: „Presne! Je nezmyselné hovoriť o polarizácii fotónu predtým, než ju nameriame.“

Študent: „Ale táto pravdepodobnosť sa náhle zmení...“

Simplicio: „*Je nezmyselné hovoriť o nej predtým, než ju nameriame!*“

Čo týmto vlastne Simplicio *myslí*? Nehľadiac na dôveryhodnosť jeho slov; aký druh stavu skutočnosti by zodpovedal tomu, že jeho slová sú pravdivé?

Akým spôsobom by skutočnosť mohla *byť*, aby bolo nezmyselné hovoriť o tom, že sa porušila špeciálna relavitita, pretože ovplyvnená vlastnosť neexistuje, hoci je možné vypočítať jej zmenu?

Ale viete čo? Kašlite na to. Chcem vedieť odpoveď na ešte dôležitejšiu otázku:

Odkiaľ Simplicio *berie* všetky tieto veci?

Predpokladajme, že vezmete Schrödingerovu rovnicu a poviete, ako pozitívny fakt:

„Táto rovnica vytvára dobré predpovede, ale nič neznamená!“

Naozaj? *Ako to viete?*

Občas chodím okolo a hovorím, že základnou otázkou rozumnosti je *Prečo veríte tomu, čomu veríte?*

Hovoríte, že Schrödingerova rovnica „nič neznamená“. Ako sa táto položka definitívneho poznania dostala do vášho vlastníctva, ak to nebola jednoducho nevedomosť nesprávne interpretovaná ako vedomosť?

Povedal vám to nejaký experiment? Som otvorený myšlienke, že experimenty nám môžu povedať veci, ktoré vyzerajú filozoficky nemožné. Ale v tomto prípade by som rád videl jasné údaje. Bol niekedy okamih, keď ste starostlivo nastavili experimentálny aparát a zistili ste, že by ste mali očakávať výsledok (1) ak Schrödingerova rovnica niečo znamená, alebo (2) ak Schrödingerova rovnica nič neznamená; a potom ste dostali výsledok (2)?

Salviati: „Ak zmeriam polarizáciu fotónu 90°, a potom zmeriam polarizáciu 45°, a potom opäť zmeriam polarizáciu 90°, moje experimentálne záznamy ukazujú, že zo 100 pokusov bol fotón pohltený 47-krát a prenesený 53-krát.“

> Simplicio: „Polarizácia 90° a polarizácia 45° sú nezlučiteľné vlastnosti; nemôžu obe *existovať* zároveň, a ak odmeriate jednu, je nezmyselné *hovoriť* o druhej.“

Ako to viete?

Ako si získal tento kúsok poznania, Simplicio? Viem, odkiaľ to *svoje* získal Salviati – ale odkiaľ pochádza to *tvoje*?

Môj postoj k otázkam existencie a významu je pekne znázornený v diskusii o súčasnom stave indície pre to, <u>či je vesmír konečný v priestore alebo nekonečný v priestore</u>, kde James D. Miller <u>napomenul Robina Hansona</u>:

> Robin, trpíš prehnaným sebavedomím, keď predpokladáš, že vesmír vôbec existuje. Iste je tu nejaká šanca, že veľkosť vesmíru je nula.

Na čo som <u>odpovedal</u>:

> James, aj keby vesmír neexistoval, aj tak by bolo pekné vedieť, či to je nekonečný alebo konečný vesmír, čo neexistuje.

Ha! Myslíte si, že ten starý trik „vesmír neexistuje“ ma zastaví? Ani ma len nespomalí!

Nie je to tak, že by som *vylučoval* možnosť, že vesmír neexistuje. Ja len toľko, že *aj keby* nič neexistovalo, stále by som chcel o tom ničom vedieť, koľko len môžem. Moja <u>zvedavosť</u> len tak nepominie kvôli tomu, že skutočnosť neexistuje, pochopte!

Podstata „skutočnosti“ je niečo, čo ma doteraz mätie, čo ponecháva otvorenú aj tú možnosť, že nič také neexistuje. Ale stále platí Eganov zákon: „To všetko dokopy dáva normálnosť.“ Jablká neprestali padať, keď Einstein vyvrátil Newtonovu teóriu gravitácie.

Iste, keď sa prach usadí, možno sa ukáže, že jablká neexistujú, Zem neexistuje, skutočnosť neexistuje. Ale tie neexistujúce jablká budú aj tak padať na neexistujúcu zem s nezmyselným zrýchlením 9,8 m/s^2.

Hovoríte, že vesmír neexistuje? Dobre, povedzme, že vám to verím – hoci mi nie je jasné, *čo* presne mám veriť, okrem toho, že <u>tie slová zopakujem</u>.

A teraz, čo sa stane, ak stlačím *toto* tlačidlo?

V kapitole <u>Jednoduchá pravda</u> som povedal:

> Úprimne, nie som si celkom istý, odkiaľ pochádza táto „skutočnosť“. Neviem si vytvoriť v laboratóriu svoju vlastnú skutočnosť, takže tomu zrejme zatiaľ nerozumiem. Ale občas sa stáva, že silno verím, že sa niečo stane, a potom sa stane niečo iné... Potrebujem teda rôzne názvy pre veci, ktoré určujú moje predpovede, a pre veci, ktoré určujú moje experimentálne výsledky. To prvé nazývam „názor“ a to druhé nazývam „skutočnosť“.

Chcete povedať, že kvantovo mechanické rovnice „nie sú skutočné“? Budem zhovievavý a budem predpokladať, že to niečo znamená. Čo by to mohlo znamenať?

Možno to znamená, že rovnice, ktoré určujú moje predpovede sú podstatne odlišné od tej veci, ktorá určuje moje experimentálne výsledky. Čo potom *určuje* moje experimentálne výsledky? Ak mi poviete, že „nič“, rád by som vedel, akého druhu je toto „nič“, a prečo toto „nič“ zdanlivo vykazuje takú pravidelnosť v určovaní napríklad mojich experimentálnych meraní hmotnosti elektrónu.

Nevychádzam dobre s ľuďmi, ktorí mi hovoria, aby som sa <u>prestal pýtať</u>. Ak mi poviete, že niečo je jednoznačne pozitívne nezmyselné, chcem vedieť, čo presne tým myslíte, a ako ste to zistili. V opačnom prípade ste mi nedali odpoveď, iba ste mi povedali, aby som sa prestal pýtať.

„Jednoduchá pravda“ opisuje život bača a jeho pomocníka, ktorí objavili, ako počítať ovce pomocou hádzania kameňov do vedier, keď ich navštívi vyslanec dvora, ktorý chce vedieť, ako tieto „čarovné kamene“ fungujú. Bača sa pokúša vysvetliť: „Prázdne vedro je magické vtedy a iba vtedy, keď na lúke nie sú ovce,“ ale čoskoro je prevalcovaný vzrušujúcou diskusiou medzi pomocníkom a vyslancom o tom, ako sa mágia mohla dostať do kameňa.

Tu máme kvantové rovnice, ktoré dávajú vynikajúce experimentálne predpovede. Čo *presne* znamená, že sú „bez významu“? Je to ako vedro s kameňmi, ktoré *funguje na počítanie oviec*, ale *nie je v ňom žiadne kúzlo*?

Predtým než Bellova veta vyvrátila lokálne skryté premenné, zdalo sa možné, že (ako si myslel Einstein) existuje nejaký úplnejší popis skutočnosti, ktorý nemáme, a že kvantová teória zhŕňa neúplné poznanie tohto úplného popisu. Zákony, ktoré sme zistili, by sa ukázali byť niečo ako zákony štatistickej mechaniky: kvantitatívne tvrdenia o neistote. To by stále nerobilo tieto rovnice „nezmyselnými“; čiastočne poznanie *je* zmysel pravdepodobnosti.

Ale Bellova veta robí omnoho menej uveriteľným, že kvantové rovnice sú čiastočným poznaním niečoho deterministického, tak ako je štatistická mechanika v klasickej fyzike čiastočným poznaním niečoho deterministického. A ešte ani vtedy by kvantové vysvetlenia neboli „nezmyselné“ v zvyčajnom zmysle tohto slova; boli by „štatistické“, „približné“, „s čiastočnou informáciou“, alebo v najhoršom prípade „nesprávne“.

Máme tu rovnice, ktoré nám dávajú vynikajúce predpovede. Hovoríte, že „nič neznamenajú“. Pýtam sa, čo je potom to, čo určuje moje experimentálne výsledky. Neviete odpovedať. Dobre, tak na základe čoho vylučujete možnosť, že nám kvantové rovnice dávajú takéto vynikajúce predpovede preto, lebo, povedzme, niečo „znamenajú“?

Nechcem tu trivializovať otázky o skutočnosti alebo o význame. Ale povedať o niečom, že to „nič neznamená“, a povedať, že tým sa debata vyriešila, ukončila, bodka, hotovo, na to musíte mať teóriu ako presne to nič neznamená. A keď *odpoviete* na toto, tá otázka by viac nemala vyzerať tajomne.

Ako si možno spomínate zo sémantických stopiek, existujú slová a slovné spojenia, ktoré ani nie sú *odpoveďami* na otázky, ako skôr poznávacími dopravnými značkami, ktoré naznačujú, že sa máte *prestať pýtať*. „Prečo vôbec niečo existuje? Lebo Boh!“ je klasický príklad, ale existujú aj iné, napríklad „Lebo élan vital!“

Povedzte ľuďom, aby „držali hubu a počítali“, pretože neviete, čo tie výpočty znamenajú, a o pár rokov sa bude „Držte hubu!“ predkladať ako pozitívna teória kvantovej mechaniky.

Mám *najvyššiu* úctu voči každému historickému fyzikovi, ktorý sa čo len *priblížil* k tomu, aby *naozaj* držal hubu a počítal; ktorý bol naozaj konzervatívny pri odhadovaní toho, čo vie a čo nevie. To je to najlepšie, čo kto mohol urobiť, ak nebol Hugh Everett, a prideľujem za to päťdesiat bodov racionality. Moje pohŕdanie je vyhradené tým, ktorí si „nevieme, prečo to funguje“ vysvetlili ako pozitívne poznanie, že tieto rovnice definitívne nie sú skutočné.

Chcem tým povedať, keby takýto trik fungoval, bol by príliš dobrý na to, aby sa obmedzil na jednu podoblasť. Prečo by fyzici nemohli využívať výhovorku „nie naozaj“ aj *mimo* kvantovej mechaniky?

> „Hej, neporušuje tvoja nová ‚teória klbka‘ špeciálnu relativitu?“
>
> „Nie, tie rovnice nič neznamenajú. Počuj, neporušuje tvoj model ‚chaoticky zlej inflácie‘ súmernosť CPT?“
>
> „Moje rovnice *znamenajú ešte menej* než tvoje rovnice! Takže tvoja kritika sa *dvojnásobne* nepočíta.“

A keby to nefungovalo, skúste si napísať vlastnú kartu „Priepustka z väzenia“.

Ak má celý tento príbeh nejaké morálne ponaučenie, je to ponaučenie, aké veľmi ťažké je zotrvať v stave *priznaného zmätku*, bez vymyslenia si príbehu, ktorý vám dá uzáver – aké ťažké je vyhnúť sa manipulovaniu s vlastnou nevedomosťou, ako keby to bola definitívna vedomosť, ktorú máte.

*→

239. Keby mnohé svety prišli ako prvé

Netvrdím, že by som to zvládol lepšie, keby som sa bol narodil v tej dobe namiesto dnešnej...

→ http://lesswrong.com/lw/q5/quantum_nonrealism/

Makroskopickú dekoherenciu, čiže mnohé svety, po prvýkrát navrhol v roku 1957 Hugh Everett III. Jeho článok bol ignorovaný. John Wheeler povedal Everettovi, aby šiel za Nielsom Bohrom. Bohr ho nebral vážne.

Everett zdrvený opustil akademickú fyziku, vymyslel všeobecné použitie Lagrangeových multiplikátorov v optimalizačných problémoch, a stal sa multimilionárom.

Až v roku 1970, keď Bryce DeWitt (ktorý zaviedol pojem „mnoho svetov“) napísal článok pre *Physics Today*, bolo všeobecné odvetvie prvýkrát informované o Everettových myšlienkach. Makroskopická dekoherencia si odvtedy stále získavala prívržencov, a teraz je už možno väčšinovým názorom (alebo možno nie).

Predstavme si však, že by si ľudia uvedomili dekoherenciu a makroskopickú dekoherenciu ihneď po objave previazania, v 1920-tych rokoch. A predstavme si, že by nikto nenavrhol teóriu kolapsu až do roku 1957. Strácala by teraz dekoherencia postupne popularitu, zatiaľ čo by teórie kolapsu naberali sily?

Predstavte si alternatívnu Zem, kde celkom prvý fyzik, ktorý objavil previazanosť a superpozíciu, povedal: „Pri všetkých svätých ohnivých opiciach, tam vonku sú zilióny iných Zemí!“

Počas nasledujúcich rokov bolo navrhovaných veľa hypotéz na vysvetlenie tajomných Bornových pravdepodobností. Ale nikto *zatiaľ* nenavrhol predpoklad kolapsu. Táto možnosť jednoducho nikomu nenapadla.

Jedného dňa Huve Erett prišiel do kancelárie Bielsa Nohra...

„Nerozumiem tomu,“ povedal Huve Erett, „prečo nikoho z fyzikov *nezaujíma* moja hypotéza. Nie sú azda Bornove štatistiky najväčšou záhadou v modernej kvantovej teórii?“

Biels Nohr si vzdychol. Za iných okolností by sa ani neunúval, ale niečo na tomto mladom mužovi ho presvedčilo, aby to skúsil.

„Huve,“ povie Nohr, „každý fyzik každoročne stretne tucet ľudí, ktorí si myslia, že vedia vysvetliť Bornove štatistiky. Ak pôjdete na párty a poviete niekomu, že ste fyzik, je šanca aspoň jedna k desiatim, že dostanete nové vysvetlenie Bornových štatistík. Je to jeden z najslávnejších problémov modernej vedy, a čo je horšie, je to problém, o ktorom si každý myslí, že mu rozumie. Aby si nová Bornova hypotéza získala pozornosť, musela by byť... čertovsky dobrá.“

„A *toto*,“ povie Huve, „*toto* nie je *dobré*?“

Huve ukáže na článok, ktorý priniesol Bielsovi Nohrovi. Je to krátky článok. Nadpis znie: „Riešenie Bornovho problému“. Text článku znie:

> „Keď vykonáte meranie kvantového systému, všetky časti vlnovej funkcie okrem jedného bodu zmiznú, pričom víťaz je určený nedeterministicky spôsobom určeným Bornovými štatistikami.“

„Rád by som sa absolútne ubezpečil,“ povie Nohr opatrne, „že vám rozumiem. Hovoríte, že máme vlnovú funkciu – ktorá sa vyvíja podľa Wheelerovej-DeWittovej rovnice – a zrazu celá táto vlnová funkcia, okrem jedinej časti, len tak spontánne klesne na nulovú amplitúdu. Všade naraz. To sa stane vtedy, keď hore na makroskopickej úrovni niečo ‚meriame‘.“

„Správne!“ povie Huve.

„Takže vlnová funkcia vie, kedy ju ‚meriame‘. Čo presne je ‚meranie‘? Ako vlnová funkcia vie, že tu sme? Čo sa dialo predtým než tu boli ľudia, ktorí mohli veci merať?“

„Ehm...“ Huve sa na chvíľu zamyslí. Potom sa načiahne pre článok, prešikrtne „keď vykonáte meranie kvantového systému“ a dopíše „keď sa kvantová superpozícia stane príliš veľkou.“

Huve sa rozžiarene pozerá. „Opravené!“

„Vidím,“ povedal Nohr. „A aká veľká je ‚príliš veľká‘?“

„Možno na úrovni 50 mikrónov,“ povie Huve. „Počul som, že to zatiaľ netestovali.“

Náhle do miestnosti strčí hlavu študent. „Hej, počuli ste o tom? Práve overili superpozíciu na úrovni 50 mikrónov.“

„Aha,“ povie Huve, „ehm, tak potom na nejakej inej úrovni. Na nejakej takej, aby experimentálne výsledky vychádzali správne.“

Nohr sa zamračí. „Pozrite, mladý muž, v tejto veci pravda nebude príjemná. Vypočujete si, čo si o tom myslím?“

„Áno,“ povie Huve, „ja chcem len vedieť, prečo ma fyzici nepočúvajú.“

„Dobre,“ povie Nohr. Vzdychne si. „Pozrite, keby táto vaša teória bola naozaj pravdivá – keby celé úseky vlnovej funkcie len tak v okamihu mizli – bolo by to... pozrime sa na to. Jediný zákon celej kvantovej mechaniky, ktorý je nelineárny, neunitárny, nediferencovateľný a nespojitý. Bránilo by to fyzike v lokálnom vývoju, kde sa každý kus pozerá iba na svojich bezprostredných susedov. Váš ‚kolaps‘ by bol jediným základným javom v celej fyzike, ktorý by mal preferovanú bázu a preferovaný priestor simultánnosti. Kolaps by bol jediným javom v celej fyzike, ktorý porušuje súmernosť CPT, Liouvilleovu vetu, a špeciálnu relativitu. Vo vašej pôvodnej verzii by kolaps bol aj jediným javom v celej fyzike, ktorý by bol principiálne myšlienkový. Vynechal som niečo?“

„Kolaps by bol zároveň jediným nezapríčineným javom,“ podotkol Hume. „Nie je vďaka tomu táto teória ešte nádhernejšia a úžasnejšia?“

„Myslím si, Huve,“ povedal Nohr, „že fyzici môžu vnímať výnimočnosť vašej teórie ako bod v jej neprospech.“

„Aha,“ povedal zaskočený Huve. „No, myslím, že by som vedel vyriešiť tú vec s nediferencovateľnosťou, keby som predpokladal výraz druhého rádu v...“

„Huve,“ povedal Nohr, „nemyslím si, že vám dochádza, čo tu naznačujem. Dôvod, prečo vám fyzici nevenujú pozornosť, je že vaša teória nie je fyzikálna. Je magická.“

„Ale Bornove štatistiky sú najväčšou záhadou modernej fyziky, a táto teória poskytuje mechanizmus pre Bornove štatistiky,“ namieta Huve.

„Nie, Huve, neposkytuje,“ povie Nohr unavene. „To je ako povedať, že ste ‚poskytli mechanizmus‘ pre elektromagnetizmus, keď poviete, že to malí anjeli posúvajú nabité častice okolo v súlade s Maxwellovými rovnicami. Namiesto povedania ‚tu sú Maxwellove rovnice, ktoré hovoria anjelom, kam majú posúvať elektróny‘ hovoríme iba ‚tu sú Maxwellove rovnice‘ a zostane nám tak striktne jednoduchšia teória. V našom prípade, nevieme, *prečo* sa dejú Bornove štatistiky. Ale vy ste neuviedli najmenší dôvod, prečo by váš ‚predpoklad kolapsu‘ mal eliminovať svety v súlade s Bornovými štatistikami a nie s niečím iným. Vy ani nevyužívate fakt, že kvantový vývoj je unitárny...“

„To preto, že nie je“ skočí mu do reči Huve.

„...o čom viacmenej každý vie, že to nejako musí byť kľúčom k Bornovým štatistikám. Namiesto toho iba hovoríte: ‚Tu sú Bornove štatistiky, ktoré hovoria kolapsu, ako má odstraňovať svety‘ a je striktne jednoduchšie povedať iba ‚Tu sú Bornove štatistiky‘.“

„Ale...“ povie Huve.

„*Navyše,*“ povie Nohr zvýšeným hlasom, „nedali ste žiadne zdôvodnenie, prečo by mal po kolapse zostať iba *jeden* jediný svet, alebo prečo kolaps nastane predtým, než sa *ľudia* dostanú do superpozície, čo robí vašu teóriu *naozaj podozrivou* pre moderného fyzika. Toto je rovnaký druh netestovateľnej hypotézy ako keď davy za ‚jedného Krista‘ argumentujú, že by sme mali ‚učiť kontroverziu‘, keď hovoríme stredoškolákom o iných Zemiach.“

„Ja nie som jedno-kristovec!“ namieta Huve.

„Dobre,“ povie Nohr, „tak *prečo* potom ale predpokladáte, že zostane iba jeden svet? A to nie je jediný problém vašej teórie. Ktorá časť vlnovej funkcie sa odstráni, presne? A v ktorej báze? Je jasné, že celá vlnová funkcia nie je skomprimovaná iba na deltu, inak by bežné kvantové počítače nemohli zostať v superpozícii, keď niekde inde nastane kolaps... dočerta, prestala by fungovať bežná molekulárna chémia...“

Huve rýchlo vo svojom článku vyškrtne „okrem jedného bodu“, dopíše „okrem jednej časti“ a potom povie: „Kolaps neskomprimuje vlnovú funkciu do jedného bodu. Odstráni všetky amplitúdy *okrem* jedného sveta, ale ponechá *všetky* amplitúdy *v* tomto svete.“

„Prečo?“ povie Nohr. „V princípe, keď raz predpokladáte ‚kolaps‘, potom by tento ‚kolaps‘ mal odstrániť všetky časti vlnovej funkcie, všade – prečo nechať len jeden pekný svet? *Vie* azda ten kolaps, že *tu sme my?*“

Huve povie: „Nechá jeden celý svet, pretože to je v súlade s našimi experimentmi.“

„Huve,“ povie Nohr trpezlivo, „toto sa nazýva ‚post hoc‘. Navyše, dekoherencia je spojitý proces. Keby ste rozdeľovali podľa celých mozgov, kde neuróny signalizujú odlišne, tieto časti majú takmer nulovú vzájomnú interferenciu vo vlnovej funkcii. Ale mnoho iných procesov sa do veľkej miery prekrýva. Nie je žiaden možný spôsob, ako ukázať na ‚jeden svet‘ a odstrániť všetko ostatné, bez celkom svojvoľného výberu, vrátane svojvoľného výberu bázy...“

„Ale...“ povie Huve.

„A *najmä*,“ povie Nohr, „ten *dôvod*, prečo mi nemôžete povedať, ktorá časť vlnovej funkcie zmizne, alebo čo presne sa stane, alebo čo presne to spustí, je že keby sme prijali túto vašu teóriu, bol by to *jediný neformálne špecifikovaný, kvalitatívne základný zákon* v celom učive fyziky. Čoskoro by sa žiadni dvaja fyzici nedokázali dohodnúť na presných detailoch! Prečo? Pretože by to bol *jediný základný zákon v celej modernej fyzike, ktorému by sme verili bez experimentálnych dôkazov, ktoré by mohli upresniť, ako funguje*.“

„Čo, naozaj?“ povie Huve. „Myslel som si, že vo fyzike je omnoho viac neformálnych vecí. Chcem tým povedať, nehovorili ste pred chvíľou o tom, že je nemožné určiť ‚jeden svet‘?“

„To je preto, lebo svety nie sú *základné*, Huve! Máme hromadu experimentálnych indícií podopierajúcich základný zákon, Wheelerovu-DeWittovu rovnicu, ktorú používame na opis vývoja vlnovej funkcie. Jednoducho použijeme presne tú istú rovnicu, aby sme dostali popis makroskopickej dekoherencie. Ak si odmyslíme komplikácie pri výpočte, táto rovnica by nám v princípe mohla povedať, kedy *presne* nastane makroskopická dekoherencia. Nevieme, odkiaľ pochádzajú Bornovské štatistiky, ale máme hromadu indícií, čo tieto Bornovské štatistiky *sú*. Ale keď sa vás opýtam, kedy alebo kde nastane kolaps, neviete – *pretože na určenie tohto nie sú absolútne žiadne experimentálne indície.* Huve, dokonca aj keby tento ‚predpoklad kolapsu‘ fungoval presne tak, ako hovoríte, *neexistoval by žiaden spôsob, ako by ste to mohli <u>vedieť</u>!* Prečo nie gazilión iných rovnako zázračných možností?“

Huve defenzívne zdvihne ruku. „Ja nehovorím, že by sa moja teória mala učiť na univerzitách ako akceptovaná pravda! Ja len chcem, aby sa experimentálne testovala! Je na tom niečo zlé?“

„Neurčitili ste, kedy nastane kolaps, takže nemôžem zostaviť test, ktorý by vašu teóriu falzifikoval,“ povie Nohr. „Keď sme si toto povedali, my už experimentálne hľadáme, či sa nejaká časť kvantových zákonov zmení na rastúcej makroskopickej úrovni. Aj kvôli všeobecným dôvodom, pre prípad, že by bolo niečo na 20. desatinnom mieste, čo sa ukáže iba v makroskopických systémoch, ale aj v nádeji, že objavíme niečo, čo vrhne nejaké svetlo na Bornove štatistiky. Pritom samozrejme kontrolujeme čas dekoherencie. Ale obzeráme *všeobecne*, čo by mohlo byť inak. Nikto nebude uprednostňovať váš nelineárny, neunitárny, nediferencovateľný, nelokálny, CPT-nesymetrický, nerelativistický, kurňa nekauzálny, rýchejší než svetlo, *sakra neformálny* ‚kolaps‘ pri hľadaní nápovedy. Nie pokiaľ neuvidí absolútne nezameniteľné indície a verte mi, Huve, vyžadovalo by to sakra veľa indície, aby to vyvážilo *takéto* chyby. Dokonca aj keby sa našli anomálie v časoch dekoherencie, čo si myslím, že sa stane, neznamenalo by to, že vysvetlením je ‚kolaps‘.“

„Čože?“ povie Huve. „Prečo nie?“

„Pretože musí byť miliarda iných vysvetlení, ktoré sú pravdepodobnejšie než porušenie špeciálnej relativity,“ povie Nohr. „Uvedomujete si, že keby sa toto naozaj dialo, potom by pri meraní polarizácie fotónu bol iba *jediný* výsledok? Meranie jedného fotónu v previazanom páre by ovplyvnilo iný fotón o svetelný rok ďalej. Einstein by z toho dostal infarkt.“

„To nebude *naozaj* porušovať špeciálnu relativitu,“ povie Huve. „Kolaps nastane presne takým spôsobom, aby vám zabránil niekedy *namerať* pôsobenie rýchlejšie než svetlo.“

„To nie je bod v prospech vašej teórie,“ povie Nohr. „A Einstein by aj z toho dostal infarkt.“

„Aha,“ povie Huve. „Tak povedzme, že relevantné vlastnosti danej častice *neexistujú* dokiaľ nenastane kolaps. Ak niečo neexistuje, potom ovplyvnenie toto nemôže porušiť špeciálnu relativitu...“

„Len si kopete hlbšiu jamu. Pozrite, Huve, ako všeobecný princíp platí, že skutočne *správne* teórie nevyvolávajú takýto stupeň zmätku. Ale hlavne, nemáte pre to žiadne indície. Nemáte žiaden logický

spôsob, ako vedieť, že kolaps sa deje, a nemáte žiaden dôvod v to veriť. Urobili ste chybu. Skrátka povedzte ‚joj' a posuňte sa ďalej vo svojom živote."

„Ale *možno* sa jedného dňa nájde indícia," povie Huve.

„Neviem si predstaviť, aká indícia by mohla určiť *túto konkrétnu* hypotézu jedného sveta ako vysvetlenie, ale každopádne, zatiaľ sme žiadnu takú indíciu *nenašli*," povie Nohr. „Nenašli sme nič, čo by to čo len matne naznačovalo! Nemôžete aktualizovať na základe indície, ktorá by teoreticky mohla jedného dňa prísť, ale zatiaľ neprišla! Práve teraz, dnes, neexistuje dôvod, prečo míňať vzácny čas myslením na toto, namiesto miliardy iných rovnako zázračných teórií. Neexistuje absolútne nič, čo by ospravedlňovalo vašu vieru v ‚teóriu kolapsu' *viac* než vieru, že sa raz naučíme prenášať správy rýchlejšie než svetlo využitím nekauzálnych účinkov modlitby k lietajúcemu špagetovému monštru!"

Huve sa narovná so zranenou dôstojnosťou. „Viete, aj keby moja teória bola nesprávna – a ja pripúšťam, že by mohla byť nesprávna..."

„*Ak?*" povie Nohr. „*Mohla?*"

„Ak, ako hovorím, je moja teória nesprávna," pokračuje Huve, „potom niekde tam vonku existuje iný svet, v ktorom ja som slávny fyzik, a *vy* ste osamelý vyhnanec!"

Nohr si zaborí hlavu do dlaní. „Ach, toto už nie. Nepočuli ste porekadlo: ‚Žite vo svojom vlastnom svete?' A zo všetkých ľudí práve *vy*..."

„Niekde tam vonku existuje svet, kde prevažná väčšina fyzikov verí v teóriu kolapsu, a nikto za posledných tridsať rokov ani len *nenavrhol* makroskopickú dekoherenciu!"

Nohr zdvihne hlavu a začne sa smiať.

„Čo je také smiešne?" pýta sa Huve podozrievavo.

Nohr sa len smeje hlasnejšie. „Ach, jaj! Ach, jaj! Vy si naozaj myslíte, Huve, že niekde tam vonku existuje svet, kde tridsať rokov vedeli o kvantovej fyzike, a nikomu ani *nenapadlo*, že by mohol byť viac než jeden svet?"

„Áno," povie Huve, „presne to si myslím."

„Ach jaj! Takže vy hovoríte, Huve, že fyzici zistili superpozíciu v mikroskopických systémoch, a vypracovali kvantitatívne rovnice, ktorými sa riadi superpozícia v každom jednom prípade, ktorý testovali. A celých tridsať rokov *jediný človek* nepovedal: ‚Hej, ktovie či tieto zákony nie sú univerzálne'."

„Prečo by to hovoril?" povie Huve. „Fyzikálne modely sa niekedy ukážu nesprávne, keď skúmate nové situácie."

„Ale že by na to ani *nepomyslel*?" povie Nohr neveriacky. „Vidíte, ako padajú jablká, vypracujete zákon gravitácie pre všetky planéty slnečnej sústavy okrem Jupitera, a ani vám *nenapadne* použiť to aj na Jupiter, pretože Jupiter je príliš veľký? To je ako nejaká komédia, kde chlapík otvorí krabicu, tam naňho vystrelí torta na pružine, tak chlapík otvorí ďalšiu krabicu, a tá obsahuje ďalšiu tortu na pružine, a chlapík v tom pokračuje, pretože ani *nepomyslí* na možnosť, že by tá ďalšia krabica tiež mohla obsahovať tortu. Myslíte si, že John von Neumann, ktorý mal asi najvyššie IQ v histórii, by na to nepomyslel?"

„Presne tak," povie Huve. „Nepomyslel. Uvažujte o tom."

„To by bol svet, kde môj dobrý priateľ Ernest sformuluje svoj myšlienkový experiment o Schrödingerovej mačke a v tomto svete ten myšlienkový experiment znie: ‚Hej, predstavme si, že máme rádioaktívnu časticu, ktorá sa dostane do superpozície rozpadu a nerozpadu. Potom táto častica interaguje so senzorom, a ten senzor sa dostane do superpozície spustenia a nespustenia. Ten senzor interaguje s výbušninou, ktorá sa dostane do superpozície vybuchnutia a nevybuchnutia; čo interaguje s mačkou, takže tá mačka sa dostane do superpozície života a smrti. Potom sa človek pozrie na túto mačku,' a v tomto bode sa Shcrödinger zastaví a povie: ‚fíha, no neviem si predstaviť čo by sa mohlo stať potom.' Takže Schrödinger to každému ukáže a všetci na to: ‚Wow, vôbec netuším, čo by sa v tomto bode mohlo stať, to je úžasný paradox.' Dokiaľ o tom konečne nepočujete *vy* a nepoviete: ‚Hej, možno v *tomto* bode polovica superpozície len tak zmizne, náhodne, rýchlejšie než svetlo' a každý na to: ‚Wow, to je super nápad!'"

„Presne tak," povie Huve. „Niekde sa toto muselo stať."

„Huve, to je svet, kde každý jeden fyzik a pravdepodobne celá poondená ľudská rasa sú príliš blbí na to, aby sa prihlásili na kryoniku! Tu hovoríme o Zemi, kde je George W. Bush prezidentom!“

*→

240. Kde sa filozofia stretáva s vedou

Keď sa obzrieme späť na raných kvantových fyzikov – nie za účelom karhať hlavné postavy, ani tvrdiť, že by sme to boli urobili lepšie, keby sme sa narodili do tej doby; ale aby sme si skúsili odniesť ponaučenie a nabudúce to robiť lepšie – keď sa obzrieme späť na temný vek kvantových fyzikov, ja by som ako „najzákladnejšiu“ chybu navrhol...

...*nie* to, že sa pokúsili obrátiť tok posledných troch tisícročí vedy, ktoré naznačovali, že myseľ je zložená vo fyzike, a nie je základom fyziky. Toto je Veda, a mávame v nej revolúcie. Z času na čas musíte zvrátiť trend. Budúcnosť je vždy absurdná, ale nikdy neporušuje zákony.

Navrhol by som, ako základnú chybu, ktorú by sme nabudúce nemali opakovať, že raní vedci zabudli, že *oni sami* sa skladajú z častíc.

Chcem tým povedať, väčšina z nich to určite v teoretickej rovine vedela.

A predsa si nevšimli, že vloženie senzora na zaznamenanie prechádzajúce elektrónu, alebo dokonca len *poznanie* histórie toho elektrónu, bolo príkladom „častíc na iných miestach“. Nevšimli si teda, že kvantová teória rôznych konfigurácií už vysvetlila experimentálny výsledok, a nebolo treba odvolávať sa na vedomie.

V pravekom prostredí ľudia často čelili adaptívne dôležitej úlohe predpovedania druhých ľudí. Za týmto účelom ste mysleli na svojich blížnych ľudí ako na niekoho, kto má myšlienky, pozná veci a cíti veci, a nie ako na niekoho, kto sa skladá z častíc. V skutočnosti mnohé kmene lovcov a zberačov možno ani nevedeli, že častice existujú. Je omnoho *intuitívnejšie – pripadá* to jednoduchšie – myslieť na to, že niekto niečo „vie“, než myslieť na to, že častice jeho mozgu sa nachádzajú v inom stave. Je ľahšie vyjadrovať svoja očakávania pomocou toho, čo *ľudia vedia*; zdá sa to tak prirodzenejšie; rýchlejšie vám to naskočí v mysli.

Rovnako ako bolo kedysi jednoduchšie predstaviť si Thora vrhajúceho blesky než predstaviť si Maxwellove rovnice – aj keď sa Maxwellove rovnice dajú opísať počítačovým programom omnoho menším ako program pre inteligentného činiteľa ako je Thor.

Preto starovekým fyzikom pripadalo prirodzené uvažovať: „Viem, kde ten fotón bol... aký rozdiel môže *toto* spôsobiť?“ Nie: „súčasný stav častíc môjho mozgu koreluje s históriou fotónu... aký rozdiel môže *toto* spôsobiť?“

A podobne, pretože sa zdalo ľahké a intuitívne modelovať skutočnosť pomocou toho, čo ľudia vedia, zatiaľ čo rozoberanie poznania na stavy mozgu neprichádza tak rýchlo na myseľ, *zdalo* sa ako jednoduchšia teória povedať, že konfigurácia môže mať amplitúdu iba „ak neviete viac“.

Aby sme z dualistickej kvantovej hypotézy urobili *formálnu* teóriu – ktorá by sa dala zapísať ako počítačový program, aby nemuseli ľudskí vedci rozhodovať, či nastalo „pozorovanie“ - museli by ste špecifikovať, čo to znamená, že „pozorovateľ“ niečo „vie“, pomocou pojmov, ktoré by váš počítač vedel spracovať.

Bude teda vaša teória fundamentálnej fyziky skúmať všetky častice v ľudskom mozgu a rozhodovať sa, či tieto častice niečo „vedia“, aby spočítala pohyby častíc? Ale ako potom spočítate pohyby častíc v samotnom mozgu? Nebola by tam možnosť nekonečnej rekurzie?

Dokiaľ však pojmy v teórii spracovávajú ľudskí vedci, *jednoducho vedeli*, kedy nastalo „pozorovanie“. Povedali ste, že „pozorovanie“ nastalo vždy vtedy, keď muselo nastať, aby experimentálne predpovede vyšli správne – jemná forma ustavičného dolaďovania.

(Pamätajte, že základy kvantovej teórie boli sformulované predtým než Alan Turing povedal čokoľvek o Turingových strojoch, a *dávno* predtým než bol všeobecne známy pojem počítačového

→ http://lesswrong.com/lw/q7/if_manyworlds_had_come_first/

výpočtu. Rozdiel medzi efektívnou formálnou teóriou a takou, ktorá si vyžaduje ľudskú interpretáciu, nebol vtedy taký jasný ako dnes. Je ľahké upresniť tieto problémy v spätnom pohľade; nemali by ste si odniesť ponaučenie, že problémy sú v spätnom pohľade zvyčajne takéto zrejmé.)

Keď sa obzrieme, môže sa to *zdať* ako jedna meta-lekcia na naučenie sa z histórie, že filozofia je vo vede naozaj dôležitá – nie je to iba nejaké pridané samostatné akademické odvetvie.

Napokon, tí raní kvantoví vedci robili všetky správne experimenty. To ich interpretácie boli mimo. A problémy interpretácií neboli dôsledkom toho, že by im zle vyšla štatistika.

Keď sa obzrieme, zdá sa, že chyby, ktoré urobili, boli chybami v tom druhu myslenia, ktoré by sme označili ako, no, „filozofické“.

Keď sa obzrieme a opýtame sa: „Ako to mohli tí raní kvantoví fyzici urobiť lepšie, aspoň principiálne?“ zdá sa, že vhľady, ktoré im chýbali, boli filozofické.

A predsa tam neboli profesionálni filozofi, ktorí by nabehli, vyriešili by problém, vyjasnili by to tajomstvo, a urobili všetko opäť normálnym. Boli to, no, fyzici.

Dá sa argumentovať, že Leibniz mal prinajmenšom predtuchu o kvantovej fyzike, ako sa kedysi myslelo, že Demokritus mal predtuchu o atómoch. Ale to je spätný pohľad. Je to výsledkom pozerania sa na výsledok, spomínania, a povedania: „Hej, Leibniz povedal niečo podobné.“

Aj keď to nejaký filozof povie správne vopred, zvyčajne je to veda, ktorá nakoniec *povie nám*, ktorý filozof mal pravdu – nie *predchádzajúci konsenzus* filozofickej komunity.

Myslím si, že to hovorí niečo zásadné o podstate filozofie a o rozhraní medzi filozofiou a vedou.

Kedysi sa hovorilo, že každá veda začína ako filozofia, ale potom vyrastie a opustí maternicu filozofie, takže v ľubovoľnom čase je „Filozofia“ to, čo sme zatiaľ nepremenili na vedu.

Tvrdím, že keď sa pozrieme na históriu kvantovej fyziky a povieme: „Potrebovali filozofické vhľady“, čo tam *naozaj* vidíme, je že ten potrebný vhľad bol v tvare, ktorý sa zatiaľ ešte neučí v štandardných akademických triedach, a ešte sa nezredukoval na výpočet.

Kedysi samotný pojem vedeckej metódy – aktualizácia názorov na základe experimentálnych indícií – bol filozofickým pojmom. Ale nepresadili ho profesionálni filozofi. Bola to moc vedy v skutočnom svete, ktorá ukázala, že vedecká epistemológia je dobrá epistemológia, nie predchádzajúci konsenzus filozofov.

Dnes sa táto filozofia aktualizovania názorov *začína* redukovať na výpočet – štatistika, bayesovská teória pravdepodobnosti.

Ale vtedy za Galileových čias to boli iba *hmlisté slovné argumenty*, ktoré hovorili, že by ste sa mali pokúšať tvoriť číselné predpovede experimentálnych výsledkov, namiesto štúdia Biblie alebo Aristotela.

Na hraniciach vedy a najmä na hraniciach vedeckého *chaosu* a vedeckého *zmätku* nájdete problémy myslenia, ktoré sa neučia v akademických kurzoch, a ktoré neboli zredukované na výpočet. A toto bude vyzerať ako oblasť filozofie; bude sa zdať, že musíte myslieť filozoficky, aby ste sa vymotali z tohto zmätku. Ale keď sa história obzrie naspäť, obávam sa, že to zvyčajne nie je profesionálny filozof, kto vyhrá všetky guličky – pretože toto filozofické myslenie si vyžaduje dôvernú účasť v danej vedeckej oblasti. Dokonca aj keď sa to dodatočne všetko zdá poznateľné a priori; a dokonca aj keď to dodatočne nejaký filozof naozaj *dostal* a priori; ešte aj tak si vyžaduje dôvernú účasť vidieť to v praxi, a experimentálne výsledky aby ste povedali svetu, ktorý filozof vyhral.

Tvrdím, že tak ako etika, aj filozofia je naozaj dôležitá, ale účinne sa praktizuje iba *zvnútra* vedy. Pokúšať sa robiť filozofiu hraníc vedy, ako samostatnú vednú disciplínu, je rovnaká chyba ako snažiť sa mať samostatných etikov. Skončí to etikmi, ktorí sa budú rozprávať najmä s inými etikmi, a filozofmi, ktorí sa budú rozprávať najmä s inými filozofmi.

Tým nechcem povedať, že na svete nie je miesto pre profesionálnych filozofov. Niektoré problémy sú také chaotické, že v komnatách vedy pre ne nie je vyhradené vôbec žiadne miesto. Ale títo „profesionálni filozofi“ by urobili veľmi, veľmi múdro, keby sa naučili každý kúsok poznania v relevantne vyzerajúcich vedách, ku ktorému sa dokážu dostať. Nemali by byť prekvapení z predstavy, že experiment a nie debata nakoniec rozhodne argumentáciu. Nemali by cúvnuť pred vykonávaním svojich vlastných experimentov, ak si nejaké dokážu vymyslieť.

Myslím si, že toto je ponaučenie z histórie.

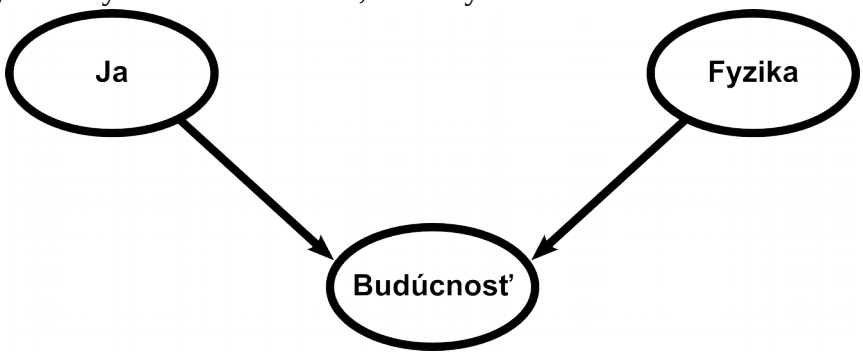
*→

241. Ty si fyzika

Pred troma mesiacmi – ježkove oči, to bolo naozaj tak dávno? - som zadal nasledujúcu domácu úlohu: Urobte výpis zásobníka ľudského poznávacieho algoritmu, ktorý vytvára debaty o „slobodnej vôli“. Všimnite si, že táto úloha je silno odlišná od debaty, či slobodná vôľa existuje alebo neexistuje.

Ako sa dalo čakať, ľudia sa pýtajú: „Ak je budúcnosť určená, ako ju naše rozhodnutia môžu ovplyvňovať?“ Múdry čitateľ dokáže uhádnuť, že to všetko dáva dokopy normálnosť; ale necháva to otázku, *ako*.

Ľudia počujú: „Vesmír funguje ako mechanické hodiny; fyzika je deterministická; budúcnosť je pevne daná.“ A ich myseľ si vytvorí kauzálnu sieť, ktorá vyzerá asi takto:



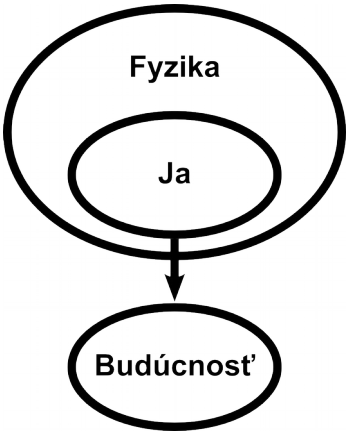
Tu vidíme, ako príčiny „Ja“ a „Fyzika“ zápasia o určenie stavu následku „Budúcnosť“. Ak je „Budúcnosť“ naozaj celá určená „Fyzikou“, potom zrejme nie je priestor, aby bola určená „Mnou“.

Táto kauzálna sieť nie je explicitný filozofický názor. Je implicitná – reprezentácia na pozadí mozgu ovládajúca, ktoré filozofické argumenty vyzerajú „rozumne“. Skrátka to vyzerá, že takto to *je*.

Z času na čas sa objaví ďalšia tlačová správa neurovedcov tvrdiacich, že, keďže výskumníci použili fMRI a zistili, že mozog robí to alebo ono počas procesu rozhodovania, *nie ste to vy, kto sa rozhoduje, je to váš mozog*.

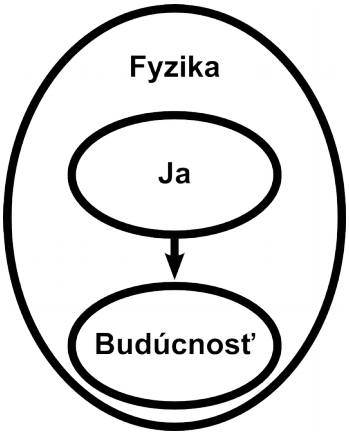
Podobne tá stará klasika: „Redukcionizmus podkopáva samotnú rozumnosť. Pretože potom vždy keby ste niečo povedali, nebol by to výsledok *rozmýšľania* o indíciách – bolo by to iba odrážanie sa kvarkov.“

Samozrejme, skutočný diagram by bol:



Alebo ešte lepšie:

→ http://lesswrong.com/lw/pg/where_philosophy_meets_science/



Prečo to nie je jasné? Pretože existujú mnohé úrovne organizácie, ktoré oddeľujú naše modely našich myšlienok – našich emócií, našich názorov, našej bolestivej nerozhodnosti, a našich konečných rozhodnutí – od našich modelov elektrónov a kvarkov.

Vieme si *intuitívne* predstaviť, že ruka sa skladá z prstov (a dlane). Pýtať sa, či je to *naozaj* naša ruka, čo niečo dvíha, alebo *iba* naše prsty a dlaň, je jasne nesprávna otázka.

Lenže medzera medzi fyzikou a poznávaním sa nedá prekročiť priamou predstavou. Nikto si nedokáže *predstaviť*, ako sa z atómov skladá človek, tak ako vidí, že sa z prstov skladá ruka.

A tak treba *neustálu bdelosť*, aby ste si udržali predstavu seba ako bytosti *v rámci fyziky*.

Táto pozornosť je jedným z veľkých kľúčov k filozofii, podobne ako Klam projekcie mysle. Spomeniete si, že toto je miesto, ktoré som nominoval ako to, kde sa kvantoví fyzici potkli, keď si nedokázali predstaviť makroskopickú dekoherenciu; nenapadlo im aplikovať tieto zákony na *seba samých*.

Názory, túžby, emócie, morálka, ciele, predstavy, očakávania, zmyslové vnemy, prchavé želania, ideály, pokušenia... Toto môžeme označiť ako „povrchovú vrstvu“ mysle, tie časti seba, ktoré ľudia vidia aj bez vedy. Keď poviem: „Nie si to *ty*, kto určuje budúcnosť, sú to tvoje *túžby, plány a činy*, ktoré určujú budúcnosť,“ rýchlo vidíte vzťah medzi celkom a časťami. Je to ihneď viditeľné, ako prsty, z ktorých sa skladá ruka. Existujú ďalšie vzťahy medzi celkom a časťami, až celkom naspodok po fyziku, ale tie nie sú ihneď viditeľné.

„Kompatibilizmus“ je filozofický názor, že „slobodnú vôľu“ možno intuitívne a uspokojujúco definovať tak, aby bola kompatibilná s deterministickou fyzikou. „Nekompatibilizmus“ je názor, že slobodná vôľa a determinizmus sú nezlučiteľné.

Môj názor by sa možno dal nazvať „Potrebizmus“. Keď sa aktivita, voľba, kontrola a morálna zodpovednosť definujú zmysluplným spôsobom, *potrebujú* determinizmus – prinajmenšom nejaké kúsky determinizmu vo vesmíre. Ak si vyberáte, plánujete, konáte, a privediete tak do existencie nejakú budúcnosť, v súlade s vašou túžbou, toto celé si vyžaduje skutočnosť s nejakým typom zákonov; nemôžete to urobiť uprostred naprostého chaosu. Musí existovať poriadok aspoň v tých častiach skutočnosti, ktoré ovládate. *Vy* ste vnútri fyziky, takže vy/fyzika ste určili budúcnosť. Keby ste neboli určovaní fyzikou, nemohla by byť určovaná vami.

Alebo by ste azda mohli povedať: „Keby budúcnosť nebola určená skutočnosťou, nemohla by byť určená vami“ alebo „Keby budúcnosť nebola niečím určená, nemohla by byť určená vami“. Nepotrebujete neurovedu ani fyziku, aby ste dostačili naivné definície slobodnej vôle do nezmyselnosti. Keby myseľ nebola stelesnená v mozgu, bola by stelesnená v niečom inom; existovala by nejaká *skutočná vec*, ktorá by bola myseľ. Keby budúcnosť nebola určená fyzikou, bola by určovaná *niečím*, nejakým zákonom, nejakým poriadkom, nejakou širšou skutočnosťou, ktorá by zahŕňala aj vás.

Ale ak nás zákony fyziky ovládajú, ako potom môžeme povedať, že ovládame sami seba?

Otočte to naopak: Keby nás zákony fyziky *neovládali*, ako by sme vôbec mohli ovládať sami seba?

Ako by myšlienky mohli posudzovať iné myšlienky, ako by emócie mohli navzájom zápasiť, ako by jeden smer aktivity mohol vyzerať najlepší, ako by sme sa mohli vo svojich plánoch presúvať od neistoty k istote, uprostred číreho chaosu?

Keby sme neboli v skutočnosti, kde by sme boli?

Budúcnosť určuje fyzika. Aká fyzika? Taká fyzika, ktoré zahŕňa akcie ľudských bytostí.

Rozhodnutia ľudí sú určené fyzikou. Akou fyzikou? Takou fyzikou, ktorá zahŕňa zvažovanie, rozhodovanie, uvažovanie o rôznych výsledkoch, ich hodnotenie, pokušenie, riadenie sa morálkou, racionalizovanie priestupkov, snahu robiť veci lepšie...

Nie je tam žiadna časť, kde priletí kvark z Pluta a prehlasuje toto všetko.

Myšlienky vo vašom rozhodovacom procese sú *skutočné*, všetky sú *niečo*. Ale myšlienka je príliš veľká a zložitá na to, aby to bol atóm. Preto sa myšlienky skladajú z menších vecí, a naše označenie pre veci, z ktorých sa skladajú veci, je „fyzika“.

Fyzika je základom našich rozhodnutí a zahŕňa naše rozhodnutia, a *nevyvracia* ich.

*→

242. Mnoho svetov, jeden najlepší odhad

Keď sa pozriete na mnohé mikroskopické fyzikálne javy – fotón, elektrón, atóm vodíka, laser – a milión iných známych experimentálnych nastavení – je možné nájsť jednoduché fyzikálne zákony, ktoré vyzerajú, že riadia všetky malé veci (dokiaľ sa nepýtate na gravitáciu). Tieto zákony riadia vývoj vysoko abstraktného a matematického objektu, ktorý som nazýval „distribúcia amplitúdy“, ale ktorý sa všeobecne označuje ako „vlnová funkcia“.

Teraz sú tu drsné otázky ohľadom správneho zovšeobecnenia, ktorým sa riadia všetky tieto drobné prípady. Nazvime objekt „zeldrý“, ak vyzerá zelený pred 1. januárom 2020, a potom vždy vyzerá modrý. Ak všetky smaragdy, ktoré sme doteraz videli, vyzerali zelené, je správne zovšeobecnenie „smaragdy sú zelené“ alebo „smaragdy sú zeldré“?

Odpoveď je, že správne zovšeobecnenie je „smaragdy sú zelené“. Nebudem teraz vysvetľovať, prečo. Nie je to témou tejto kapitoly, a samozrejmá odpoveď v tomto prípade je zhodou okolností tá správna. Skutočná Cesta nie je hlúpa: nech ste so svojou logikou akokoľvek chytrí, nakoniec by mala prísť k správnej odpovedi a nie k nesprávnej.

V podobnom zmysle, *najjednoduchšie* zovšeobecnenia, ktoré by mohli pokryť pozorované *mikroskopické* javy, majú tvar „všetky elektróny majú spin 1/2“ a nie „všetky elektróny majú spin 1/2 pred 1. januárom 2020“ alebo „všetky elektróny majú spin 1/2, pokiaľ nie sú súčasťou previazaného systému, ktorý váži viac ako 1 gram“.

Keď obrátime svoju pozornosť na makroskopické javy, náš pohľad je zakrytý. Nemôžeme experimentovať s vlnovou funkciou človeka rovnakým spôsobom, ako môžeme experimentovať s vlnovou funkciou atómu vodíka. V žiadnom prípade nedokážete naozaj prečítať vlnovú funkciu pomocou malého kvantového skenera. Ale v prípade povedzme človeka, veľkosť celého organizmu presahuje našu schopnosť robiť presné výpočty alebo presné pokusy – nemôžeme potvrdiť, či sa kvantové rovnice dodržiavajú *do posledného detailu*.

Vieme, že javy, ktoré bežne označujeme ako „kvantové“, len tak nezmiznú, keď sa nazbiera veľa mikroskopických predmetov. Laser vysiela záplavu koherentných fotónov, namiesto aby robil povedzme niečo celkom iné. Atómy majú také chemické vlastnostni, ktoré by mali mať podľa kvantovej teórie, čo im umožňuje spájať sa do stabilných molekúl, z ktorých sa skladá človek.

Takže v istom zmysle máme hromadu indícií, že kvantové zákony sa spájajú do makroskopickej úrovne bez významenejšieho rozdielu. Chémia funguje aj vo veľkom.

Nemôžeme si však priamo overiť, že častice, z ktorých sa skladá človek, majú agregátnu vlnovú funkciu, ktorá sa správa *presne* tak, ako hovoria najjednoduchšie kvantové zákony. Ach, vieme, že molekuly a atómy nezmiznú, vieme, že makroskopické zrkadlá stále odrážajú pod uhlom dopadu. Môžeme dostať *veľa* predpovedí na vysokej úrovni vychádzajúc z predpokladov, že mikroskopické aj makroskopické sa riadi tými istými zákonmi, a každá testovaná predpoveď vyšla rovnako.

→ http://lesswrong.com/lw/r0/thou_art_physics/

Ale keby niekto tvrdil, že makroskopické kvantový obraz sa líši od mikroskopického v nejakom zatiaľ ešte netestovateľnom detaile – niečo, čo sa ukáže iba na nemerateľnom 20. desatinnom mieste mikroskopických interakcií, ale agreguje sa do niečoho väčšieho pri makroskopických interakciách – nuž, nemôžeme *dokázať*, že sa mýli. To iba Occamova britva hovorí: „Existujú zilióny nových základných zákonov, ktoré by si mohol predpokladať pre 20. desatinné miesto; prečo vôbec *myslíš* na tento konkrétny?“

Ak počítame pomocou najjednoduchších zákonov, ktorými sa riadia všetky známe prípady, zistíme, že ľudia skončia v stave kvantovej superpozície, rovnako ako fotóny v superpozícii odrazenia sa od polopriepustného zrkadla a prejdenia cezeň. V situácii Schrödingerovej mačky ide nestabilný atóm do superpozície rozpadu a nerozpadu. Senzor, nastavený na tento atóm, ide do superpozície spustenia a nespustenia. (V skutočnosti je tá superpozícia teraz spoločným stavom [atóm-rozpadnutý × senzor-spustený] + [atóm-stabilný × senzor-nespustený].) Výbušnina pripojená na senzor ide do superpozície vybuchnutia a nevybuchnutia; mačka v krabici ide do superpozície smrti a života; a človek, ktorý pozrie do krabice, ide do superpozície pozvracania sa a pokoja. Ten istý zákon na každej úrovni.

Ľudia, ktorí interagujú so systémami v superpozícii, sa sami vyvinú do superpozície. Ale mozog, ktorý vidí vybuchnutú mačku, a mozog, ktorý vidí živú mačku, majú mnoho odlišne vysielajúcich neurónov, čiže *mnoho* častíc v iných polohách. Sú v konfiguračnom priestore veľmi vzdialené a komunikujú v exponenciálne mikroskopickej miere. Nie na 30. desatinnom mieste, ale na 10^{30}. desatinnom mieste. Žiadna konkrétna myseľ, žiaden konkrétny poznávací kauzálny proces, nevidí rozmazanú superpozíciu mačiek.

Fakt, že *vy* vidíte iba živú mačku, *alebo* iba mŕtvu mačku, je presne to, čo by predpovedali tie najjednoduchšie kvantové zákony. Nemáme teda dôvod veriť, podľa našej doterajšej skúsenosti, že kvantové zákony sa nejako odlišujú na makroskopickej a mikroskopickej úrovni.

A fyzici overili superpozíciu na čoraz väčších úrovniach. Zdá sa, že sa práve usilujú testovať superpozíciu na 50-mikrónových predmetoch, čo je väčšie než väčšina neurónov.

Existencia iných verzií nás samotných, vlastne iných Zemí, nie je predpokladaná *dodatočne*. Jednoducho predpokladáme, že tie isté zákony riadia všetky úrovne, lebo nemáme dôvod predpokladať inak, a všetky experimentálne testy zatiaľ uspeli. Existencie inej dekoherentnej Zeme je *logickým dôsledkom* toho najjednoduchšieho zovšeobecnenia, ktoré zodpovedá všetkým známym faktom. Ak si myslíte, že Occamova britva hovorí, že iné svety sú „nepotrebné veci“ navyše, potom by ste si mali skontrolovať matematiku teórie pravdepodobnosti, pretože takto Occamova britva nefunguje.

Pri snahe rozšíriť mikroskopické zákony všeobecne, vrátane supoerpozície ľudí, je stále jedna konkrétna záhada, ktorá vyzerá čudne:

Ak skúšame dostať pravdepodobnosti pomocou počítania rozdielnych pozorovateľov, nie je žiaden *zrejmý* dôvod, prečo by integrál druhej mocniny absolútnej hodnoty vlnovej funkcie mal korelovať so štatistickými výsledkami experimentu. Nie je žiaden známy dôvod pre Bornove pravdepodobnosti, a dokonca sa zdá, že by sme a priori mali očakávať pravdepodobnosť 50/50, že ľubovoľný binárny kvantový pokus pôjde oboma smermi, ak budeme iba počítať pozorovateľov.

Robin Hanson navrhuje, že keby exponenciálne podpriemerne malé dekoherentné škvrny amplitúdy („svety“) interferovali s exponenciálne malými únikmi z väčších škvŕn, dostali by sme z toho Bornove pravdepodobnosti. Považujem to za zaujímavú možnosť, pretože je to také normálne.

(Osobne som nedávno uvažoval iným smerom: Ak sa pokúšam počítať pozorovateľov samozrejmým spôsobom, dostanem vo všeobecnosti zvláštne vyzerajúce výsledky, nie iba v prípade kvantovej fyziky. Keby som napríklad skopíroval svoj mozog na bilión podobných verzií, podľa toho, či som vyhral lotériu, kým som v anestézii; dovolil by som svojim ja, aby sa prebudili a prípadne sa od seba navzájom trochu odlišovali; a potom by som ich opäť spojil do jedného ja; potom počítanie pozorovateľov zvyčajným spôsobom hovorí, že by som mal dokázať spôsobiť, že vyhrám lotériu (ak dokážem svoj mozog kopírovať a spájať, čo by myseľ nahraná v počítači mohla dokázať).

V tejto súvislosti mi príde veľmi zaujímavé, že Bornovo pravidlo *nemá* problém s kopírovaním a opätovným spájaním. Ak je daná unitárna kvantová fyzika, Bornovo pravidlo je *jediné* pravidlo, ktoré bráni „pozorovateľom“ získať nadprirodzené schopnosti – čo *nevysvetľuje* Bornovo pravidlo, ale je to iste *zaujímavý fakt*. Ak je dané Bornovo pravidlo, dokonca aj rozdeľovanie a spájanie svetov by stále viedlo ku konzistentným pravdepodobnostiam. Možno fyzika používa lepší antropický princíp než ja!

Možno by som sa mohol poučiť od fyziky namiesto snahy vymyslieť to a priori, a pozrieť, kam ma to zavedie? Ale *zatiaľ* ma to nezaviedlo nikam, takže toto je sotva „odpoveď“.)

Wallace, Deutsch a ďalší sa snažia odvodiť Bornovo pravidlo z teórie rozhodovania. Voči tomuto som dosť podozrievavý, lebo sa mi zdá, že existuje zložka „Čo sa mi stane?“, ktorú nemôžem zmeniť pomocou zmeny svojej funkcie úžitku. Aj keby som sa absolútne *nezaujímal* o svety, v ktorých nevyhrám kvantovú lotériu, stále mi pripadá, že je nejaký zmysel, v ktorom sa prebudím „prevažne“ v svetoch, kde som lotériu nevyhral. A o tomto si myslím, že to treba vysvetliť.

Pointa je, že bolo navrhnutých veľa hypotéz ohľadom Bornových pravdepodobností. Nie toľko, koľko by malo byť, pretože toto tajomstvo sa už dlho neprávom označuje ako „vyriešené“. Ale aj tak, návrhov bolo veľa.

Je legitímna nádej, že sa Bornova záhada vyrieši bez nových základných zákonov. Váš svet sa nerozdeľuje na presne dva nové podprocesy v presne tom okamihu, keď uvidíte „POHLTENÝ“ alebo „PRENESENÝ“ na LCD monitore svojho senzora fotónu. Stále sme v superpozícii a dekoherencii, nepretržite, občas v spojitých rozmeroch – hoci mozgy sú digitálne a zahŕňajú signály celých neurónov, a signál/nesignál by bol extrémne dekoherovaný stav čo len pre *jediný* neurón... Zdalo by sa, že je tu miesto na *niečo* nečakané, čo môže za Bornove štatistiky – lepšie pochopenie antropickej váhy pozorovateľov, alebo lepšie pochopenie superpozícií mozgu – bez nových základov.

Nemôžeme však vylúčiť možnosť, že v Bornových štatistikách je zahrnutý nový základný zákon.

Ako povedal Jess Riedel:

> Ak si z dejín fyziky môžeme vziať nejaké ponaučenie, je to, že vždy, keď sme skúšali novú pokusnú „situáciu“ (napr. veľkú rýchlosť, malú veľkosť, veľkú hustotu, veľkú energiu), pozorovali sme javy, ktoré viedli k novým teóriám (v danom poradí: špeciálna relativita, kvantová mechanika, všeobecná relativita, a štandardný model).

„Vždy“ je príliš silné. Technický detail, iste, ale dôležitý: nemôžete len tak *predpokladať*, že nejaký konkrétny zákon v novej situácii zlyhá. Ale je možné, že v Bornových štatistikách je zahrnutý nejaký nový základný zákon, ktorý sa na mikroskopickej úrovni prejavuje iba na 20. desatinnom mieste (a preto nebol zatiaľ zistiteľný), ale jeho účinky sa nasčítajú na dôležité na makroskopickej úrovni.

Mohol by existovať nejaký zákon, zatiaľ ešte neobjavený, ktorý spôsobuje, že existuje iba *jeden* svet?

To je šokujúca predstava; vyplývalo by z toho, že všetky naše dvojčatá v iných svetoch – všetky rôzne verzie nás samotných, ktoré sa neustále oddeľujú, nie iba keď výskumníci robia kvantové merania, ale aj bežnými entropickými procesmi – sú naozaj *preč*, a zostali sme sami! Táto verzia Zeme by bola *jediná* verzia, ktoré existuje v lokálnom priestore! Keby sa inflačný scenár v kozmológii ukázal ako nesprávny, a topológia vesmíru by bola zároveň konečná a relatívne malá – takže by Zem nemala vzdialené duplikáty, ktoré by vyplývali z exponenciálne rozľahlého vesmíru – potom by táto Zem mohla byť jedinou Zemou, ktorá *kdekoľvek* existuje, čo je pomerne znervózňujúca myšlienka!

Ale je nebezpečné sústrediť sa natoľko na konkrétnu hypotézu, na ktorú nemáte žiaden konkrétny dôvod myslieť. To je tá istá základná chyba ako u ľudí od Inteligentného Dizajnu, ktorí si vezmú ľubovoľnú záhadu v modernej genetike a povedia: „Vidíte, toto musel urobiť Boh!“ Prečo „Boh“ a nie zilión iných možných vysvetlení, na ktoré ste mali pomyslieť dávno predtým než ste vyslovili hypotézu nadprirodzeného zásahu, ak nie preto, lebo ste už potajomky začali tak, že ste už poznali odpoveď, ku ktorej ste chceli dôjsť.

Nemali by ste sa ani *pýtať*: „Môže existovať iba jeden svet?“, ale namiesto toho ísť robiť fyziku, a baviť sa o tejto *konkrétnej* téme iba ak si to bude vyžadovať nová indícia.

Môže existovať nejaký, zatiaľ ešte neznámy základný zákon, ktorý dáva vesmíru privilegovaný stred, ktorý sa zhodou okolností nachádza v našej Zemi – čo by dokazovalo, že Kopernik sa celý čas mýlil a Biblia mala pravdu?

Ak sa pýtate *túto konkrétnu* otázku – a nie zilión iných otázok, kde stredom vesmíru je Proxima Centauri, prípadne sa ukáže, že vesmír má obľúbený druh pizze a je to feferónková – prezrádzate tým svoj skrytý zámer. A hoci si to neosvietený človek nemusí uvedomiť, dávať vesmíru privilegovaný stred, *ktorý nasleduje Zem putujúcu priestorom*, by bolo pomerne zložité na *matematicky jednoduchý* základný zákon.

Podobne je to s pýtaním sa, či by mohol existovať iba jeden svet. Prezrádza to sentimentálnu pripútanosť k ľudskej intuícii, ktorá sa už ukázala ako nesprávna. Koleso vedy sa otáča, ale neotáča sa *naspäť*.

Máme konkrétne dôvody vysoko podozrievať predstavu iba jediného sveta. Pojem „jeden svet“ existuje na vyššej úrovni organizácie, ako poloha Zeme v priestore; na kvantovej úrovni nie sú pevné hranice (hoci mozgy, ktoré sa odlišujú signálmi celých neurónov, sú iste dekoherované). Ako by *základný* fyzikálny zákon identifikoval jeden svet *na vyššej úrovni*?

Omnoho horšie, ľubovoľný scenár, v ktorom by prežil iba *jediný* svet, takže ľubovoľné meranie by malo iba *jeden* výsledok, by porušoval špeciálnu relativitu.

Ak sú isté zákony pravdivé na všetkých úrovniach – čiže ak sú správne mnohé svety – potom keď meriate jeden z dvojice previazaných polarizovaných fotónov, skončíte vo svete, v ktorom je tento fotón polarizovaný povedzme zvislo, a vaše alternatívne verzie skončia vo svetoch, kde je fotón polarizovaný vodorovne. Z vášho pohľadu pred meraním je pravdepodobnosť 50/50. O niekoľko svetelných rokov ďalej niekto odmeria ten druhý fotón v uhle 20° voči vašej báze. Z ich pohľadu je tiež pravdepodobnosť dostania ľubovoľného okamžitého výsledku 50/50 – udržiavajú si nemenný stav zovšeobecneného previazania s vašou vzdialenou polohou, bez ohľadu na to, čo robíte. Ale keď sa o niekoľko rokov neskôr obaja stretnete, vaša pravdepodobnosť stretnúť sa s priateľom, ktorý dostal *rovnaký* výsledok, je 11,6 % a nie 50 %.

Keby existoval iba jeden globálny svet, potom by existoval iba jeden výsledok ľubovoľného kvantového merania. Buď by ste namerali fotón polarizovaný zvislo alebo vodorovne, ale nie oboje. O niekoľko svetelných rokov ďalej by sa pravdepodobnosť, že niekto iný nameria podobne polarizovaný fotón v báze otočenej o 20° naozaj *zmenila* z 50/50 na 11,6 %.

Nemôžete toto interpretovať ako prípad púheho odhalenia vlastností, ktoré tam už boli; toto vylučuje Bellova veta. Zdá sa, že neexistuje žiaden konzistentný pohľad na vesmír, kde obe kvantové merania majú jediný výsledok a pritom sú obe merania vopred určené, žiadne z nich neovplyvňuje to druhé. Niečo sa musí naozaj *zmeniť*, rýchlejšie než svetlo.

A toto sa zdá byť plne všeobecnou námietkou nielen voči teóriám kolapsu, ale voči ľobovoľnej teórii, ktorá nám dáva jeden globálny svet! Neexistuje konzistentný pohľad, v ktorom merania majú jeden výsledok, ale sú lokálne určené (hoci aj lokálne náhodne určené). Nejaké tajomné vplyvy musia preskočiť priestorovú vzdialenosť.

Toto nie je triviálna vec. Nemôžete sa zachrániť tým, že zamávate rukami a poviete: „vplyv cestuje naspäť v čase do vytvorenia previazaných fotónov, potom dopredu v čase k druhému fotónu, takže v skutočnosti nikdy nepreskakuje priestorovú vzdialenosť“. (Tento pohľad bol naozaj predložený, čo vám dáva nejakú predstavu o veľkosti paradoxu vyplývajúceho z jedného globálneho sveta!) Jedno meranie muselo zmeniť druhé, takže ktoré meranie bolo *prvé*? Existuje globálny priestor súčasnosti? Nemôžete mať obe merania ako „prvé“, pretože podľa Bellovej vety neexistuje spôsob, ako by lokálna informácia mohla zodpovedať za pozorované výsledky, atď.

Zhodou okolností sa tento experiment už urobil, a keby existoval nejaký tajomný vplyv, musel by cestovať šesť miliónkrát rýchlejšie než svetlo v referenčnej sústave Švajčiarskych Álp. Ďalej, experimentálne sa ukázalo, že tomuto tajomnému vplyvu nezáleží na tom, či tieto dva fotóny meriame v referenčnom rámci, ktorý by spôsobil, že jedno meranie nastane „pred druhým“.

Špeciálna relativita pripadá nám ľuďom kontraintuitívna – ako svojvoľné ohraničenie rýchlosti, ktoré by ste dokázali obísť, keby ste sa vrátili naspäť v čase, a potom zase dopredu. Zákon, pri ktorom by ste sa vedeli vyhnúť trestu za porušenie, keby sa vám podarilo skryť svoj zločin pred úradmi.

Lenže špeciálna relativita v skutočnosti hovorí, že ľudské intuície ohľadom priestoru a času sú jednoducho nesprávne. Krížom cez priepasti priestoru *neexistuje* žiadne globálne „teraz“, *neexistuje* „predtým“ ani „potom“. Schopnosť *predstaviť si* jeden globálny svet, *čo len v princípe*, znamená, že na intuitívnej úrovni nerozumiete špeciálnej relativite. V opačnom prípade by bolo zrejmé, že fyzika sa šíri lokálne s invariantnými stavmi vzdialeného previazania, a že informácie potrebné na podporu *jedného globálneho sveta* jednoducho *lokálne neexistujú*.

Možno má táto napohľad bezchybná logika nejakú chybu – možno moje použitie Bellovej vety a relativity na vylúčenie jedného globálneho sveta obsahuje nejaký skrytý predpoklad, ktorý si neuvedomujem...

...ale pomyslite na to, aké bremeno teraz leží na pleciach teórie jedného sveta! *V prvom rade* nie je absolútne žiaden dôvod mať podozrenie na jeden globálny svet; to skrátka *nie je to, čo hovorí moderná fyzika!* Jeden globálny svet je stará ľudská intuícia, ktorá bola *vyvrátená*, podobne ako predstava univerzálneho absolútneho času. Princíp superpozície vidno aj v polopriepustných zrkadlách; experimenty ho overujú na čoraz väčších úrovniach superpozície – ale najmä *už neexistuje žiaden dôvod privilegovať hypotézu* jedného globálneho sveta. Spod tejto ľudskej intuície bol už odkopnutý rebrík.

Neexistujú experimentálne indície, že makroskopický svet je jeden (o mikroskopickom svete už vieme, že je v superpozícii). A táto predstava nevyhnutne porušuje buď špeciálnu relativitu, alebo si vyžaduje ešte zázračnejšie vyzerajúci skok a porušuje napohľad bezchybnú logiku. To druhé je samozrejme v praxi omnoho uveriteľnejšie. Ale nie je to až *také* uveriteľné v absolútnom zmysle. Je to vo všeobecnosti *zlé znamenie*, keď musíte predpokladať svojvoľné logické zázraky *bez experimentálnej indície*.

Čo sa týka kvantového nerealizmu, pripadá mi ako nič iné než karta Priepustka-z-väzenia. „Je v poriadku porušovať špeciálnu relativitu, pretože nič z tohto nie je naozaj!“ Nedá sa rozumne predpokladať, že tieto rovnice dávajú také vynikajúce predpovede *doslova bezdôvodne*. Bellova veta vylučuje samozrejmú možnosť, že kvantová teória predstavuje nedokonalé poznanie niečoho lokálne nedeterministického.

Navyše, makroskopická dekoherencia nám dáva dokonale *realistické* pochopenie toho, čo sa deje, kde rovnice dávajú také dobré predpovede preto, lebo odrážajú skutočnosť. A tak predstava, že kvantové rovnice „nič neznamenajú“ a preto je okej, ak porušujú špeciálnu relativitu, takže predsa len môžeme mať jeden globálny svet, je *nepotrebná*. Mne pripadá kvantový nerealizmus ako veľký bluf postavený okolo sémantických stopiek ako „Nezmyselné!“

Nie je celkom bezpečné povedať, že existencia viacerých Zemí je rovnako dobre podložená ako ľubovoľná iná vedecká pravda. Existencia iných kvantových svetov nie je rovnako dobre podložená ako existencia stromov, ktoré si väčšina z nás môže osobne pozrieť.

Možno je niečo na tom 20. desatinnom mieste, čo sa pri makroskopických udalostiach nasčíta na niečo väčšie. Možno existuje medzera v napohľad oceľovej logike, ktorá hovorí, že ľubovoľný jeden globálny svet musí porušovať špeciálnu relativitu, pretože informácia na podporu jedného globálneho sveta lokálne neexistuje. Možno sa s nami lietajúce špagetové monštrum iba zahráva a svet, ako ho poznáme, je lož.

Takže jediné, čo môžeme povedať o existencii viacerých Zemí je, že je to rovnako rozumne pravdepodobné ako napríklad tvrdenie, že rotujúce čierne diery neporušujú zákon zachovania momentu rotácie. Máme extrémne základné dôvody, ktoré súvisia s rotačnou súmernosťou priestoru, na podozrenie, že zachovanie momentu rotácie je zabudované do základnej podstaty fyziky. A nemáme žiaden konkrétny na podozrenie, že pri vysokých energiách nastane toto *konkrétne* porušenie našich starých zovšeobecnení.

Ale zatiaľ sme naozaj neskontrolovali zachovanie momentu rotácie pri rotujúcej čiernej diere – pokiaľ viem. (A keďže teraz hovorím o rozumných odhadoch v stavoch čiastočného poznania, pointa je úplne rovnaká, ak sa toto pozorovanie už urobila, akurát ja o ňom ešte neviem.) A čierna diera je omnoho

masívnejšia situácia. Takže poslušnosť čiernych dier nie je *až taká* bezpečná ako to, že môj záchod pri splachovaní zachováva moment rotácie, hoci keď sa nad tým zamyslím, ani to som ešte neskontroloval...

Ak však predsa urobíte tú *chybu*, že budete príliš silno rozmýšľať nad touto jednou konkrétnou možnosťou namiesto ziliónov iných možností – a najmä ak nerozumiete základnému dôvodu, *prečo* sa moment rotácie zachováva – potom vám môže začať pripadať stále uveriteľnejšie, že „rotujúca čierna diera porušuje zachovanie momentu rotácie“, ako myslíte na stále hmlistejšie uveriteľne znejúce dôvody, prečo by to *mohla* byť pravda.

Ale rozumná pravdepodobnosť je čertovsky malá.

Podobne aj rozumná pravdepodobnosť, že existuje iba jedna Zem.

Spomínam to, aby som vysvetlil môj zvyk rozprávať, akoby mnohé svety boli samozremý fakt. Mnohé svety *sú* samozrejmý fakt, ak máte všetky predpoklady pekne zoradené (rozumiete tej najzákladnejšej kvantovej fyzike, poznáte formálnu teóriu pravdepodobnosti Occamovej britvy, rozumiete špeciálnej relativite, atď.). Je mi v skutočnosti výrazne *viac* samozrejmé než predpoklad, že rotujúca čierna diera musí zachovávať moment rotácie.

Jediný dôvod, prečo sa mnohé svety všeobecne neuznávajú ako priama predpoveď fyziky, ktorej porušenie by si vyžadovalo zázrak, je tá zhoda okolností v dejinách vedy na našej Zemi, ktorá na akademickej pôde zabetónovala teóriu typu flogiston, v ktorej nepozorovateľný magický „kolaps“ rýchlejší než svetlo požiera všetky ostatné svety. A mnoho akademických fyzikov nemá matematické chápanie Occamovej britvy, čo je zvyčajná metóda na vyháňanie neviditeľných anjelov z fyziky. Takže keď natrafia na mnohé svety a je to v rozpore s ich (podkopanou) intuíciou, že existuje iba jeden svet, povedia „Ach, to je pridávanie vecí“ - čo je z hľadiska teórie pravdepodobnosti vyslovená chyba – a pokračujú vo svojom každodennom živote.

Nie som akademik. Nemusím sa klaňať nejakému služobne staršiemu fyzikovi, ktorý nepochopil niečo samozrejmé, ale ktorý bude písať recenzie na moje články do časopisov. Nemusím sa báť, že mi bude odmietnutá definitíva na základe toho, že straším svojich žiakov „sci-fi príbehmi o iných Zemiach“. Ak *ja* nemôžem hovoriť na rovinu, kto už môže?

Dovoľte mi teda povedať, veľmi jasne, v mene všetkých fyzikov okolo, ktorí sa to neodvažujú povedať sami: Mnohé svety *jednoznačne vyhrávajú* pri našom dnešnom stave indícií. Neexistuje dôvod predpokladať jednu Zem, o nič viac než predpokladať, že zrážka dvoch horných kvarkov spôsobí rozpad porušujúci zákon zachovania energie. Vyžadovalo by si to viac než nejaký neznámy základný zákon; vyžadovalo by si to zázrak.

Táto debata by už mala byť za nami. Mala by byť za nami už vyše päťdesiat rokov. Stav indícií je príliš jednostranný na to, aby ospravedlňoval ďalšiu argumentáciu. V tejto téme neexistuje rovnováha. Nie je tu žiadna rozumná kontroverzia, ktorú by sme mohli učiť. Zákony pravdepodobnosti sú zákony, nie odporúčania; nie je tu pružnosť ohľadom najlepšieho odhadu pri týchto indíciách. Naše deti sa obzrú na fakt, že začiatkom 21. storočia sme sa o tomto EŠTE STÁLE HÁDALI, a správne si odvodia, že sme boli blázni.

Robíme našej Zemi dosť dlho hanbu tým, že nedokážeme vidieť zrejmé veci. Preto v mene cti mojej Zeme píšem o existencii mnohých svetov ako o dokázanom fakte, pretože to tak *je*. Jediná otázka je, ako dlho bude ľuďom v tomto svete trvať, než aktualizujú.

*→

→ http://lesswrong.com/lw/q8/many_worlds_one_best_guess/

T: Veda a rozumnosť

243. Zlyhania vedy predkov

Tentokrát neboli žiadne rúcha, žiadne kapucne, žiadne masky. Očakávalo sa, že sa študenti stali priateľmi a spojencami. A každý vedel, prečo ste v tejto triede. Bolo by márne predstierať, že nie ste v Konšpirácii.

Ich *sensei* bol Jeffreyssai, ktorý bol možno tým najlepším zo svojich čias, za svojich čias. Jeho študentmi boli buď tí najsľubnejší žiaci, alebo tí, v ktorých výchove videli *beisutsukai* politickú výhodu.

Brennan patril do tej druhej kategórie, a vedel o tom. Ani neváhal používať meno svojej Milenky na otváranie dverí. Pri hľadaní poznania ste použili všetky dostupné cestičky; tu sa to rešpektovalo.

„...vyše tridsať rokov," povedal Jeffreyssai. „Ani jeden z nich to nevidel; ani Einstein, ani Schrödinger, dokonca ani von Neumann." Otočil sa od svojho skicára, smerom k triede. „Pýtam sa vás: Ako to, že zlyhali?"

Študenti si vymenili letmé pohľady, prepočet vzájomného rizika medzi tými opatrnými a tými, ktorý boli iba zaskočení. Vedelo sa, že Jeffreyssai hráva hry.

Nakoniec sa Hiriwa zvaná Čierna naklonila dopredu, zľahka hrkotajúc, ako sa jej náramky s vytesanými rovnicami posúvali na členkoch. „Podľa rokov, ktoré ste uviedli, *sensei*, to bolo dvesto päťdesiat rokov po Newtonovi. Vedci tej doby iste museli intuitívne chápať pojem všeobecného zákona."

„Poznať všeobecný zákon gravitácie," povedal študent Taji, sediaci neďaleko, „nie je to isté ako rozumieť *pojmu* všeobecný zákon." Bol to jeden z tých sľubných, rovnako ako Hiriwa.

Hiriwa sa zamračila. „Nie... hovorilo sa, že Newton bol uctievaný *za* objav prvého všeobecného zákona. Dokonca už vo svojej dobe. Takže sa o tom vedelo." Hiriwa sa zastavila. „Lenže sám Newton už vtedy nebol. Bolo snáď nejaké *náboženské* tabu proti navrhovaniu ďalších všeobecných zákonov? Zdráhali sa z úcty voči Newtonovi, alebo čakali, kedy prehovorí jeho *duch*? Neviem, čo motivovalo vedu predkov..."

„Nie," zamrmlal Taji so smiechom v hlase, „naozaj, *naozaj* nevieš."

Jeffreysaiov výraz bol láskavý. „Hiriwa, nebolo to náboženstvo, ani to nebolo olovo v pitnej vode, ani nemali všetci Alzheimera, ani celý deň neposedávali a nečítali internetové komixy. Zabudni na katalóg hrôz dávnych čias. Uvažuj v pojmoch kognitívnych chýb. Čo si mohla veda predkov *myslieť* nesprávne?"

Hiriwa si s povzdychom sadla. „*Sensei*, naozaj si nedokážem predstaviť prúser, ktorý by spôsobil *toto*."

„Nebola by to iba *jedna* chyba," opravil ju Taji. „Ako sa hovorí: Chyby necestujú sami; lovia vo svorkách."

„Ale *celá* ľudská rasa?" povedala Hiriwa. „Tridsať *rokov*?"

„Nebola to celá ľudská rasa, Hiriwa," povedal Styrlyn. Bol to jeden zo staršie vyzerajúcich študentov, mal krátku briadku so sivými škvrnami. „Možno jeden zo stotisíc by dokázal spamäti napísať Schrödingerovu rovnicu. Takže to mohla byť ich prvá a základná chyba – neschopnosť sústrediť svoje sily."

„*Ušetri nás propagandy!*" Jefreyssaiov pohľad bol náhle divoký. „Nie si tu na propagáciu Kooperatívnej Konšpirácie, vážený pán politik! Neohýbaj pravdu, aby si mal názor! Pokiaľ viem, vaša Konšpirácia má heslo: ‚Komparatívna výhoda.' *Naozaj* si myslíš, že by bolo pomohlo, keby celá ľudská rasa, ako v tej dobe existovala, išla debatovať o kvantovej fyzike?"

Styrlyn ani okom nemihol. „Asi nie, *sensei*," povedal. „Ale ak porovnáme našu dobu s ich dobou, stojí to za úvahu."

Jeffreyssai mával rukou rovno vo vzduchu; gesto „možno", ktoré používal na odmietnutie argumentu, ktorý bol pravdivý ale nesúvisel s témou. „Nie je to to, čo by som označil za *základnú* chybu. Vyriešenie tohto problému by si nemalo vyžadovať miliardu fyzikov."

„Viem si predstaviť *konkrétnejšie* dávne hrôzy," povedal Taji. „Tráviť celý deň písaním grantových žiadostí. Učiť vysokoškolákov, ktorí by radšej boli niekde inde. Musieť uverejniť tridsať článkov ročne, aby človek dostal definitívu..."

„Ale my sa teraz nebavíme iba o vedcoch s nižším spoločenským postavením," povedala Yin; na tvári mala mierne provokujúci úškrn. „O Schrödingerovi sa hovorí, že sa na mesiac utiahol do vily, kde mu jeho milenka poskytovala inšpiráciu, a vrátil sa so svojou slávnou rovnicou. Považujeme to za slávny historický úspech našej metodológie. Niektorí dávni fyzici *chápali*, ako sústrediť svoju myšlienkovú energiu; a mohli mať na to dosť vysoké postavenie, ak chceli."

„To je pravda," povedal Taji. „Napokon, administratívne bremená sú iba všeobecné prekážky. Podobne ako odpovede typu: ‚Nemali výcvik v teórii pravdepodobnosti, a nepoznali kognitívne skreslenia.' Náš sensei zrejme chce nejakú konkrétnejšiu odpoveď."

Jeffreyssai povzbudzujúco nadvihol obočie. „Nezavrhuj svoju líniu myšlienok tak rýchlo, Taji; začína byť k veci. Aký druh systému vytvára administratívne bremená pre svojich vlastných ľudí?"

„Systém, ktorý nedokáže primerane podporovať svojich ľudí," povedal Styrlyn. „Taký, ktorý si nedokáže vážiť ich prácu."

„Ach," povedal Jeffreyssai. „Ale je tu jeden študent, ktorý ešte nič nepovedal. *Brennan?*"

Brennan nevyskočil. Úmyselne počkal akurát tak dlho, aby ukázal, že sa nebojí, a potom povedal: „Nedostatok pragmatickej motivácie, sensei."

Jefreyssai sa zľahka usmial. „Rozveď to."

Aký druh systému by vytváral administratívne bremená pre svojich vlastných ľudí? pýtal sa ich *sensei*. Ostatní študenti sledovali svoje vlastné myšlienkové nitky. Brenan sa toho zdržal, mohol venovať viac pozornosti tým pár náznakom svojho učiteľa. Byť začiatočníkom nebola *vždy* nevýhoda – a jeho naučili, dávno predtým než ho prijali Bayesovci, využívať každú dostupnú výhodu.

„Projekt Manhattan," povedal Brennan, „bol spustený s konkrétnym *technologickým* zámerom: zbraň ohromnej moci, počas vojny. Ale chyba, ktorej sa veda predkov dopustila vzhľadom na kvantovú fyziku, nemala na ich technológiu žiaden okamžitú dopad. Boli zmätení, ale necítili zúfalú *potrebu* poznať odpoveď. Inak by systém okolo nich odstránil všetky bremená brzdiace ich úsilie vyriešiť to. Projekt Manhattan to iste tak urobil – Taji? Vieš o tom?"

Taji vyzeral zamyslene. „Nie *všetky* bremená – ale som si celkom istý, že nemuseli písať grantové požiadavky uprostred svojej práce."

„Takže," povedal Jeffreyssai. Podišiel pár krokov, stál priamo pred Brennanovou lavicou. „Ty si myslíš, že sa dávni vedci jednoducho dosť nesnažili. Pretože ich umenie nemalo žiadne vojenské využitie? Pomerne *kompetitívny* uhol pohľadu, zdá sa mi."

„Nie nevyhnutne," povedal pokojne Brennan. „Aj pragmatizmus je cnosťou rozumnosti. Túžba *používať* lepšiu kvantovú teóriu by dávnym vedcom pomohla mnohými spôsobmi, nielen motivačne. Bola by usmernila ich zvedavosť, ukázala by im, čo je úspech a čo zlyhanie."

Jeffreyssai sa zľahka zachichotal. „Nesnaž sa tak silno uhádnuť, čo chcem počuť, Kompetitor. Tvoje prvé tvrdenie bolo bližšie k môjmu skrytému cieľu; tvoja ach-aká-Bayesovská výhovorka vyšla naprázdno... Faktor, na ktorý som myslel, Brennan, bol, že dávni vedci si mysleli, že je *prijateľné*, keď im vyriešenie problému zaberie tridsať rokov. Celý ich spoločenský proces vedy bol založený na tom, že sa *jedného dňa* dostanú k pravde. Nesprávna teória sa *jedného dňa* vyradí – až vyrastie nasledujúca generácia študentov, ktorí budú poznať jej náhradu. Ako sa hovorí, práca narastá, aby vyplnila pridelený čas. Lenže ľudia dokážu myslieť na dôležité veci v kratšom čase než tridsať rokov, ak od seba *očakávajú* rýchlosť." Jeffreyssai náhle udrel dlaňou o rameno Brennanovho kresla. „*Koľko času máš na to, aby si sa vyhol letiacemu nožu?*"

„Veľmi málo času, sensei!"

„*Menej ako sekundu! Útočia na teba dvaja súperi! Koľko času máš na to, aby si uhádol, ktorý je nebezpečnejší?*"

„Menej ako sekundu, sensei!"

„*Dvaja súperi sa rozdelili a zaútočili na dve tvoje priateľky. Koľko času máš na to, aby si sa rozhodol, ktorú z nich máš naozaj rád?*“

„Menej ako sekundu, sensei!“

„*Nový argument ukazuje, že tvoja obľúbená teória je chybná! Koľko času máš na to, aby si zmenil svoj názor?*“

„Menej ako sekundu, sensei!“

„*ZLE! NEDÁVAJ MI NESPRÁVNU ODPOVEĎ IBA PRETO, LEBO ZAPADÁ DO POHODLNÉHO VZORCA A ZDÁ SA, ŽE JU OD TEBA ČAKÁM!* Ako dlho to naozaj trvá, Brennan?“

Na Brennanovom chrbte sa vytváral pot, ale on sa zastavil a naozaj sa nad tým zamyslel...

„*ODPOVEDZ, BRENNAN!*“

„*Nie, sensei! Ešte som neskončil rozmýšľanie, sensei! Odpoveď by bola predčasná! Sensei!*“

„*Výborne! Pokračuj! Ale nech ti to netrvá tridsať rokov!*“

Brennan dýchal zhlboka a usporiadaval si myšlienky. Nakoniec povedal: „Realisticky, sensei, ideálny prípad by bol taký, že by som ten problém uvidel ihneď; použil by som cvičenie na zastavenie úsudku; skúsil by som znovu zhromaždiť všetky indície skôr než by som pokračoval; a podľa toho, nakoľko by som bol citovo pripútaný k danej teórii, by som použil techniku krízy viery, aby som sa poistil, že môžem naozaj ísť ľubovoľným smerom. Takže aspoň päť minút a možno až hodinu.“

„*Dobre! Tentokrát si nad tým naozaj rozmýšľal! Rozmýšľaj nad tým vždy! Prelom stereotypy!* V časoch vedy predkov, Brennan, nebolo nezvyčajné, ak grantová agentúra strávila šesť mesiacov vyhodnocovaním žiadosti. *Dopriali si toľko času!* Teba hodnotím podľa tvojej *rýchlosti*, Brennan! Otázka neznie, či sa tam dopracuješ jedného dňa! Hocikto dokáže nájsť pravdu, ak má na to päť tisíc rokov! Musíš *ísť rýchlejšie!*“

„*Áno, sensei!*“

„Takže, Brennan, naučil si sa teraz niečo nové?“

„Áno, sensei!“

„Ako dlho ti trvalo naučiť sa túto novú vec?“

Svojvoľná voľba... „Menej ako minútu, sensei, podľa hranice, ktorá vyzerá najprirodzenejšie.“

„Menej ako minútu,“ zopakoval Jeffreyssai. „Takže, Brennan, ako dlho si myslíš, že by ti malo trvať vyriešiť väčší vedecký problém, keby si vôbec neplytval časom?“

Toto bola najzákernejšia otázka, akú Brennan kedy počul. Nedalo sa uhádnuť, aké časové obdobie má Jeffreyssai na mysli – čo by *sensei* považoval za príliš dlho alebo príliš krátko. Čo znamenalo, že jediný spôsob z kaše bol skrátka hľadať skutočnú pravdu; to by mu dalo obranu úprimnosti, aj keď to bola slabá obrana. „Jeden rok, sensei?“

„Myslíš si, že by sa to dalo urobiť za jeden mesiac, Brennan? Predpokladajme taký prípad, kde v princípe už máš dostatok experimentálnych indícií na určenie odpovede, ale nie dosť experimentálnej indície na to, aby si si mohol dovoliť robiť chyby pri jej vyhodnocovaní.“

Opäť, nedalo sa uhádnuť, akú odpoveď Jeffreyssai *chce* počuť... „Jeden mesiac mi znie ako nerealisticky krátky čas, sensei.“

„*Krátky čas?*“ povedal Jeffreyssai neveriacky. „Koľko minút má tridsať dní? Hiriwa?“

„43 200, sensei,“ odpovedala. „Ak predpokladáme pravidelný spánok a šestnásť hodín bdenia denne, potom 28 800 minút.“

„Predpokladajme, Brennan, že ti trvá celých päť minút, pomyslieť si nejakú *originálnu* myšlienku, namiesto naučenia sa od niekoho iného. Vyžaduje si hoci aj väčší vedecký problém viac ako 5760 rôznych postrehov?“

„Priznávam sa, sensei,“ povedal Brennan pomaly, „že som nad tým takýmto spôsobom nikdy nerozmýšľal... ale hovoríte, že toto je *naozaj* realistická úroveň produktivity?“

„Nie,“ povedal Jeffreyssai, „ale takisto nie je realistické myslieť si, že jeden problém vyžaduje 5760 postrehov. A áno, už sa to podarilo.“

Jeffreyssai ustúpil späť a dobrotivo sa usmial. Každý študent v miestnosti stuhol; ten úsmev už poznali. „Hoci nikto z vás nedošiel k tej konkrétnej odpovedi, na ktorú som myslel, vaše odpovede predsa boli rovnako rozumné ako tá moja. Okrem Styrlynovej, obávam sa. Dokonca ani Hiriwina odpoveď nebola celkom zlá: úloha navrhovať nové teórie sa kedysi považovala za posvätnú povinnosť vyhradenú pre tých, ktorí mali vysoké postavenie, pretože vtedy bolo v obehu obmedzené množstvo problémov. Ale *Brennanova* odpoveď je *osobitne* zaujímavá, a ja mám v úmysle otestovať jeho teóriu motivácie.“

A sakra, povedal si Brennan potichu. Jefreyssai gestom vyzval Brennana, aby sa postavil pred triedu.

Keď Brennan vstal, Jeffreyssai sa pohodlne posadil do Brennanovho kresla.

„Brennan-sensei,“ povedal Jeffreyssai, „máte päť minút na vymyslenie niečoho šokujúco skvelého, čo nám poviete o zlyhaní vedy predkov v kvantovej fyzike. A my ostatní máme zase za úlohu hľadieť na vás s očakávaním. Sotva si viem predstaviť, aké zahanbujúce by bolo, keby ste nevymysleli nič dobré.“

Hajzel. Brennan to nepovedal nahlas. Tajiho tvár ukazovala istý stupeň súcitu; Styrlyn sa tváril povznesene nad touto hrou; ale Yin naňho hľadela so sardonickým záujmom. Čo je horšie, Hiriwa naňho pozerala s očakávaním, predpokladajúc, že sa postaví voči tejto výzve. A Jeffreyssai civel s rozšírenými očami, čakajúc na guruove slová múdrosti. *Do hája, sensei.*

Brennan nepanikáril. Od najhroznejšej situácie, v akej sa kedy ocitol, to malo veľmi, veľmi, veľmi ďaleko. Chvíľku sa rozhodoval, ako bude rozmýšľať; potom rozmýšľal.

O štyri minúty a tridsať sekúnd Brennan prehovoril. (Takéto veci mali svoje umenie; keď už ich musíte urobiť tak či tak, môžete ich urobiť tak, aby vyzerali jednoducho.)

„Jedna múdra žena,“ povedal Brenann, „mi raz povedala, že je múdre pozerať na svoje staré ja ako na bláznov, pre ktorých niet spásy – aby sme vnímali ľudí, ktorými sme kedysi boli, ako úplných idiotov. Ja toto celkom netvrdím; ale toto mi povedala a je v tom viac než len zrnko pravdy. Dokiaľ robíme výhovorky pre svoju minulosť, pokúšame sa ju vykresliť lepšie, *ctíme* si ju, nedokážeme urobiť čistý zlom. Zdá sa mi, že toto pravidlo môže byť rovnaké aj pre ľudskú civilizáciu. Skúsil som sa teda obzrieť a považovať dávnych vedcov jednoducho za hlupákov.“

„Čo samozrejme neboli,“ povedal Jeffreyssai.

„Čo samozrejme neboli,“ pokračoval Brennan. „Merítkom hrubej inteligencie ma nepochybne prevyšovali. Ale napadlo mi, že problém uvidieť, čo dávni vedci robili nesprávne, môže byť v tom, že si tieto prastaré a legendárne mená príliš vážime. A to naozaj prinieslo jeden postreh.“

„Dosť bolo úvodu, Brennan,“ povedal Jeffreyssai. „Ak si našiel postreh, sem s ním.“

„Dávni vedci neboli cvičení...“ Brennan sa zastavil. „Nie, *necvičení* nie je to správne slovo. Boli cvičení na *nesprávnu úlohu*. V tej dobe neboli žiadne Konšpirácie, žiadne tajné pravdy; akonáhle dávni vedci vyriešili väčší problém, uverejnili toto riešenie pre celý svet a pre seba navzájom. Skutočne strašné a mätúce *otvorené problémy* boli extrémne nedostatkové, a spotrebovali sa v tej chvíli, keď sa vyriešili. Nebolo teda možné vycvičiť dávnych výskumníkov, *aby vo vede vytvárali poriadok z chaosu*. Boli cvičení na niečo iné – nie som si istý, čo presne...“

„Cvičení pracovať s tou vedou, ktorá *už* bola objavená,“ povedal Taji. „Pre dávnych učiteľov bolo dosť ťažké vycvičiť svojich študentov, aby *používali existujúce vedomosti*, alebo sa riadili už známymi metodológiami; to bolo všetko, čo sa dávni učitelia vedy snažili odovzdať.“

Brennan prikývol. „Čo je *veľmi* odlišná vec než sám vytvárať novú vedu. Dávni vedci, ktorí čelili problémom s kvantovou teóriou, sa možno nikdy predtým nestretli s takýmto druhom *strachu* – strachom s nepoznaného. Dávni vedci sa možno predčasne zmocnili neuspokojivých odpovedí, pretože boli zvyknutí pracovať s uhladenými, schválenými sadami poznatkov.“

„*Dobre*, Brennan,“ zamrmlal Jeffreyssai.

„Ale najmä,“ pokračoval Brennan, „si dávny vedec nemohol *cvičiť* ten skutočný problém, ktorému čelili kvantoví vedci – ako vyriešiť veľký zmätok. To bolo niečo, čo človek urobil raz za život, ak mal šťastie, a ako povedala Hiriwa, Newton tam už vtedy nebol. Takže hoci tí dávni fyzici, ktorí domotali kvantovú teóriu, neboli neinteligentní, boli to v prísnom zmysle slova *amatéri* – náhodne improvizujúci s celým procesom zmeny paradigmy.“

„A nemali teóriu pravdepodobnosti,“ poznamenala Hiriwa. „Takže ak aj niekto v tomto probléme *uspel*, netušil, čo vlastne urobil. Nedokázal to nikomu vysvetliť, nanajvýš nejasne.“

„Áno,“ povedal Styrlyn. „A bolo iba pár ľudí, ktorí sa vôbec mohli do tohto problému pustiť, či už s výcvikom alebo bez neho; to sú tí fyzici, ktorých mená dnes poznáme. Hŕstka ľudí, každý z nich urobil hŕstku objavov. Nebolo by to dosť na udržanie komunity. Každý dávny vedec čeliaci novej zmene paradigmy by musel znovu objaviť potrebné pravidlá od nuly.“

Jeffreyssai vstal spoza Brennanovej lavice. „Prijateľné, Brennan; v skutočnosti si ma prekvapil. Budem nad touto tvojou metódou musieť ešte rozmýšľať.“ Jeffreyssai prešiel k dverám triedy, potom sa obzrel. „Mal som však na mysli ešte aspoň jednu *ďalšiu* veľkú chybu vedy predkov, ktorú nikto z vás nespomenul. Očakávam, že zajtra dostanem zoznam možných chýb. Očakávam, že chyba, na ktorú myslím, bude na tom zozname. Máte na to 480 minút, keď nerátam čas na spánok. Vidím vás tu piatich. Vyriešenie tejto úlohy si nevyžaduje viac než 480 postrehov, ani viac než 96 na seba nadväzujúcich postrehov.“

A Jefreyssai opustil miestnosť.

*→

244. Dilema: veda alebo Bayes?

„Eli: V poslednej dobe píšeš veľa o fyzike. Prečo?“

--Shane Legg (a niekoľko ďalších ľudí)

> „Vo svetle tvojho vysvetlenia QM, ktoré mi znie dokonale logicky, sa zdá *samozrejmé a normálne*, že mnohé svety sú drvivo pravdepodobné. Vyzerá to takmer príliš dobré na to, aby to bola pravda, že *ja* teraz chápem niečo, čo kopa geniálnych kvantových fyzikov stále nechápe. […] Iste, vedel by som všetko nejako vyvrátiť, a aj tak si myslím, že máš pravdu, iba mi je podozrivé, že verím prvému uveriteľnému vysvetleniu, ktoré som našiel.“
>
> --Zotavujúci sa iracionalista

Zotavujúci sa iracionalista, netušíš, ako som rád, že si napísal tento komentár.

Samozrejme som mal viac než *jeden* dôvod stráviť všetok ten čas blogovaním o kvantovej fyzike. Rád mám veľa skrytých motívov, to je najbližšie, ako sa viem eticky dostať k roli superzloducha.

Ale aby som dal príklad cieľa, ktorý môžem dosiahnuť *iba* debatou o kvantovej fyzike...

Vo fyzike môžete dostať absolútne jednoznačné veci. Nie v tom zmysle, že tie veci sa jednoducho vysvetľujú. Ale ak sa pokúsite aplikovať Bayesa na zdravotnú starostlivosť, alebo na ekonomiku, možno nedokážete *formálne* vysvetliť, čo je tá najjednoduchšia hypotéza, alebo čo podporujú indície. Ale ak poviem „makroskopická dekoherencia je jednoduchšia než kolaps“, je to naozaj *striktná* jednoduchosť; mohli by ste napísať obe tieto hypotézy ako počítačové programy a porovnať počet riadkov kódu. Ani nie sú pochybnosti o samotných indíciách.

Chcel som veľmi jasný prípad – *Bayes hovorí „cik“, toto je cak* – keď nastane čas prelomiť svoju vernosť Vede.

„Ach, iste,“ poviete, „fyzici to s tými mnohými svetmi poplietli, ale daj im vydýchnuť, Eliezer! Nikto nikdy netvrdil, že spoločenský proces vedy je dokonalý. Ľudia sú nedokonalí; robia chyby.“

Ale fyzici, ktorí odmietajú prijať mnohé svety, neporušujú pravidlá Vedy. Oni *dodržiavajú* pravidlá Vedy.

Tradícia, ktorá sa odovzdávala celé generácie, hovorí, že nová fyzikálna teória prichádza s novými experimentálnymi predpoveďami, ktoré ju odlišujú od starej teórie. Vykonáte test, a nová teória sa potvrdí alebo falzifikuje. Ak sa potvrdila, usporiadate veľkú oslavu, zvoláte novinárov, a rozdáte všetkým Nobelove ceny; každý roztrasený starý profesor na dôchodku, ktorý odmietne konvertovať, bude potichu vysmievaný. Ak sa teória nepotvrdí, jej hlavný predkladateľ verejne odvolá a získa body za poctivosť.

→ http://lesswrong.com/lw/q9/the_failures_of_eld_science/

To nie je ako sa veci vo vede *robia*; ale skôr, ako by veci vo Vede *mali* fungovať. Je to ideál, o ktorý sa všetci dobrí vedci usilujú.

Teraz prídu mnohé svety, a nezdá sa, že by robili nejaké nové predpovede oproti starej teórii. To je podozrivé. A sú tam všetky tie iné svety, ale vy ich nevidíte. To je *naozaj* podozrivé. Skrátka to nevyzerá vedecky.

Ak ste sa dostali tak ďaleko ako ZsI – že vám mnohé svety pripadajú dokonale logické, samozrejmé a normálne – *a* ak ste zároveň začali ako Tradičný Racionalista, mali by ste sa byť schopný prepínať tam a naspäť medzi Vedeckým pohľadom a Bayesovským pohľadom, ako Neckerova kocka.

Nasaďte si teraz svoje Vedecké okuliare – stále ich máte niekde poruke, však? Zabudnite na všetko, čo viete o Kolmogorovskej zložitosti, Solomonoffovej indukcii alebo Minimálnej dĺžke správy. To nie je časť tradičného výcviku. Na určovanie, aké „jednoduché“ je niečo, sa používa iba približný odhad. Slovo „testovateľné“ vám nevyvolá myšlienkový obraz Bayesovej vety riadiacej toky pravdepodobnosti; vyvolá vám to myšlienkový obraz laboratória, kde sa robí experiment, a potom nasleduje oslava (alebo verejné odvolanie).

> *Nasadené Vedecké okuliare:* Súčasná kvantová teória zatiaľ splnila všetky experimentálne testy. Mnohé svety nerobia žiadne nové testovateľné predpovede – predpovedané úžasné nové javy sú všetky skryté tam, kde ich nemôžeme vidieť. Môžete ďalej fungovať aj bez predpokladu iných svetov, a presne to by ste mali urobiť. Celá táto vec zaváňa vedeckou fantastikou. Ale musíme priznať, že kvantová fyzika je veľmi hlboká a veľmi mätúca téma, a ktovie, aké objavy na nás ešte čakajú? Ozvite sa mi, keď mnohé svety urobia testovateľnú predpoveď.

Zložené Vedecké okuliare, nasadené Bayesovské okuliare:

> *Nasadené Bayesovské okuliare:* Najjednoduchšie kvantové rovnice, ktoré zahŕňajú všetky známe indície, nemajú špeciálnu výnimku pre hmotu veľkosti človeka. Neexistuje ani žiaden dôvod klásť túto konkrétnu otázku. Ďalší!

Okej, tak je toto problém, ktorý vieme opraviť za päť minút pomocou kúska lepiacej pásky a sekundového lepidla?

Nie.

He? Prečo skrátka nenaučiť nové vysokoškolské ročníky vedcov Solomonoffovu indukciu a Bayesovo pravidlo?

Pred stáročiami bola rozšírená myšlienka, že Múdri môžu odhaľovať tajomstvá vesmíru tým, že na ne budú myslieť, zatiaľ čo ísť von a *pozrieť* sa na veci bolo nižšie, podradné, naivné, a mohlo vás v konečnom dôsledku zavádzať. Nemohli ste dôverovať tomu, ako veci *vyzerajú* – iba myšlienka mohla byť vaším sprievodcom.

Veda začala ako povstanie proti tejto Hlbokej Múdrosti. V jej jadre je pragmatická viera, že ľudské bytosti sediaci v kreslách a snažiace sa o Hlbokú Múdrosť sa iba vznášajú smerom do rozprávkovej ríše. Nemôžete dôverovať svojim myšlienkam. Musíte robiť experimentálne predpovede – predpovede, ktoré pred vami ešte nikto neurobil – zbehnúť test a potvrdiť výsledok. Toto je indícia. Sedieť v kresle a myslieť na to, čo znie rozumne... by nespôsobilo *predsudky* voči vašej teórii, pretože Veda nebola idealistická viera ohľadom pragmatizmu, ani ohľadom ručnej práce. Skôr to bolo rozhodnutie, že iba experiment má posledné slovo. Iba experimenty môžu súdiť vašu teóriu – nie vaša národnosť, ani vaše náboženské presvedčenie, ani fakt, že ste túto teóriu vymysleli sediac v kresle. Iba experimenty! Ak ste sedeli v kresle a vymysleli teóriu, ktorá dávala novú predpoveď, a experiment potvrdil vašu predpoveď, potom sa budeme starať iba o výsledok tohto experimentu, a nie odkiaľ pochádza vaša hypotéza.

Toto je Veda. A ak hovoríte, že mnohé svety by mali nahradiť nesmierne úspešnú Kopenhagenskú intepretáciu, pridaním všetkých týchto dvojníkov Zeme, ktorých nemôžeme pozorovať, iba preto, lebo to *znie rozumnejšie a elegantnejšie* – nie preto, lebo to *rozdrvilo starú teóriu lepšími experimentálnymi predpoveďami* – potom rúcate základné vedecké pravidlo, ktoré bráni ľuďom pribehnúť a doplniť do všetkých teórií anjelov, pretože anjeli sú rozumnejší a elegantnejší.

Myslíte si, že naučiť pár ľudí Solomonoffovu indukciu vyrieši *tento* problém? Laureát Nobelovej ceny Robert Aumann – ktorý ako prvý dokázal, že Bayesovskí aktéri s podobnými pôvodnými pravdepodobnosťami sa nemôžu zhodnúť, že sa nezhodnú – je veriaci ortodoxný žid. Aumann pomáhal pri projekte hľadania „Biblických kódov“, skrytých proroctiev od Boha v Starom Zákone – a došiel k záveru, že projekt nepotvrdil existenciu týchto kódov. Chcete, aby si Aumann myslel, že keď raz máte Solomonoffovu indukciu, môžete zabudnúť na experimentálnu metódu? Myslíte si, že by mu to pomohlo? A väčšina vedcov sa nedostane ani na úroveň Roberta Aumanna.

Okej, opäť nasadiť Bayesovské okuliare. *Naozaj* budete veriť, že veľké časti vlnovej funkcie zmiznú, ked ich viac nevidíme. V dôsledku jediného nelineárneho, neunitárneho, nediferencovateľného, CPT-nesymetrického, nekauzálneho, rýchlejšieho než svetlo, neformálne špecifikovaného javu v celej fyzike? Jedine preto, lebo historickou zhodou okolností táto hlúpa verzia teórie bola predložená ako prvá?

Chcete urobiť veľkú zmenu vo vedeckom modeli, a veriť v zilióny iných svetov, ktoré nemôžete vidieť, bez definitívneho momentu experimentálneho triumfu nad starým modelom?

Alebo chcete odmietnuť teóriu pravdepodobnosti?

Zložíte svoju vernosť do rúk Vedy alebo Bayesa?

Michael Vassar raz skonštatoval (nie celkom vážne), že je dobrá vec, že väčšina ľudí verí v Boha, pretože inak by preňho bolo veľmi ťažké odmietnuť majoritarianizmus. Ale keďže väčšinový názor, že Boh existuje, je jednoducho neuveriteľný, nemáme na výber a musíme odmietnuť extrémne silné filozofické argumenty pre majoritarianizmus.

Môžete vidieť (jeden z dôvodov), prečo som šiel tak ďaleko vo vysvetľovaní kvantovej teórie. Tí, ktorí ste dobrí v matematike, by ste si teraz mali vedieť *predstaviť* aj makroskopickú dekoherenciu, aj teóriu pravdepodobnosti jednoduchosti a testovateľnosti – aby ste šialenstvo jedného globálneho sveta pocítili na *inštinktívnej* úrovni.

Chcel som vám predložiť jednoduchú, ostrú dilemu medzi odmietnutím vedeckej metódy alebo prijatím šialenstva.

Prečo? Dám vám nápovedu: Nie preto, lebo som zlý. Keby ste hádali moje motívy, hľadajte ďalej než iba po prvú samozrejmú odpoveď.

PS: Ak sa pokúsite prísť s nejakými chytrými spôsobmi, ako vykorčuľovať z tejto dilemy, v nasledujúcich kapitolách dostanete na zadok. Varoval som vás.

*→

245. Veda nedôveruje vašej rozumnosti

Scott Aaronson naznačuje, že mnohé svety a libertariánstvo sú si podobné v tom, že sú to oba prípady postavenia sa guľke do cesty namiesto vyhýbania sa:

> Libertariánstvo a MWI sú obe veľké filozofické teórie, ktoré začínajú z predpokladov, s ktorými súhlasia takmer všetci vzdelaní ľudia (kvantová mechanika v jednom prípade, základy ekonómie v druhom prípade) a tvrdia, že došli k záverom, ktoré väčšina vzdelaných ľudí odmieta, alebo sú z nich aspoň zmätení (existencia paralelných vesmírov, vhodnosť odstránenia požiarnych staníc).

Tak *toto* je analógia, ktorá by mne nikdy nenapadla.

Už som tvrdil, že Veda odmieta mnohé svety, ale Bayes ich prijíma. (Kde „Veda“ je s veľkým „V“, pretože hovoríme o idealizovanej forme Vedy, nie iba o skutočnom spoločenskom procese vedy.)

Ďalej sa mi zdá, že existuje *hlboká* analógia medzi libertariánstvom (s malým „l“) a Vedou:

- Obe sú založené na pragmatickej nedôvere k rozumne znejúcim argumentom.
- Obe sa snažia budovať systémy, ktoré sú dôveryhodnejšie než ľudia v nich.
- Obe prijímajú, že ľudia majú chyby, a snažia sa využiť ich chyby na poháňanie systému.

→ http://lesswrong.com/lw/qa/the_dilemma_science_or_bayes/

Ústredným argumentom libertariánstva je historicky motivovaná nedôvera k pekným teóriám „o čo *lepšia* by bola spoločnosť, keby sme zaviedli pravidlo, ktoré hovorí XYZ.“ Keby takýto trik naozaj *fungoval*, potom by viac regulácií korelovalo s vyšším ekonomickým rastom, ako by sa spoločnosť hýbala z lokálnych optím ku globálnym. Lenže keď nejaká osoba alebo záujmová skupina získa dosť moci, aby začala robiť všetko, čo považuje za dobrý nápad, história hovorí, že čo sa v skutočnosti *stane*, je revolučné Francúzsko alebo sovietske Rusko.

Plány, ktoré by podľa krásne znejúcej teórie mali urobiť každého naveky šťastným, nemajú výsledky, ktoré im predpovedali rozumne znejúce argumenty. A moc korumpuje, a priťahuje skorumpovaných.

Preto regulujete čo najmenej, pretože nemôžete dôverovať krásne znejúcim teóriám, ani ľuďom, ktorí ich implementujú.

Neukazujete prstom na ľudí, že sú sebeckí. Snažíte sa vybudovať efektívny systém produkcie zo sebeckých účastníkov vyžadovaním, aby transakcie boli dobrovoľné. Ľudia sú tak nútení hrať hry s kladným súčtom, pretože tak primäjú *druhú* stranu k podpisu zmluvy. Pri obmedzení násilia a presadzovaní zmlúv môže sebectvo jednotlivcov poháňať globálne produktívny systém.

Samozrejme, nič z tohto nefunguje v praxi rovnako dobre ako v teórii, a nebudem teraz rozoberať zlyhania trhu, problémy spoločných statkov, atď. Ústredný argument libertariánstva nie je, že libertariánstvo by fungovalo v dokonalom svete, ale že v skutočnom živote upadá elegantne. Prinajmenšom upadá menej nešikovne než ľubovoľný iný známy ekonomický princíp. (Ľudia, ktorí vidia Libertariánstvo ako dokonalé riešenie pre dokonalých ľudí, mi pripadajú, že im akosi uniká celá pointa tejto „pragmatickej nedôvery“.)

Veda pôvodne vznikla ako vzbura proti dôverovaniu Aristotelovmu slovu. Keby účastníci tejto revolúcie povedali iba: „Dôverujme sami sebe, nie Aristotelovi!“, iba by sa blysli a zhasli, podobne ako francúzska revolúcia.

Vedecká revolúcia však pretrvala, pretože – podobne ako americká revolúcia – jej architekti navrhli čudnejšiu filozofiu: „Nedôverujme nikomu! Dokonca ani sebe!“

Na začiatku prišla predstava, že nemôžeme len tak vyhodiť Aristotelove jalové špekulácie a nahradiť ich *inými* jalovými špekuláciami. Potrebujeme sa porozprávať s Prírodou a naozaj *počúvať*, čo nám odpovie. Práve toto bol geniálny nápad.

Ale potom prišli problémy s implementovaním. Ľudia sú tvrdohlaví, a nemusia byť ochotní prijať rozhodnutie experimentu. Budeme na nich kývať prstom a hovoriť „Fuj“?

Nie; budeme predpokladať a zmierime sa s tým, že každý jednotlivý vedec môže byť šialene naviazaný na svoje osobné teórie. Ani nebudeme predpokladať, že možno niekoho od týchto sklonov odučiť – nepokúsime sa vybrať Najvyšších Sudcov, ktorí budú údajne nestranní.

Namiesto toho sa pokúsime *zapriahnuť* tvrdohlavú túžbu jednotlivých vedcov dokázať svoju osobnú teóriu tým, že im povieme: „Urobte novú experimentálnu predpoveď, a urobte ten experiment. Ak máte pravdu, a ak sa experiment dá replikovať, vyhráte.“ Dokiaľ vedci veria, že toto platí, majú motív robiť experimenty, ktoré môžu *falzifikovať* ich teórie. Iba ten, kto sa zmieri s možnosťou porážky, dostane šancu vyhrať. A každé veľké tvrdenie si bude vyžadovať replikáciu; to dáva vedcom motív byť poctiví, aby sa vyhli veľkému zahanbeniu.

A tak sme tvrdohlavosť jednotlivých vedcov zapriahli do produkcie stáleho prúdu vedomostí na skupinovej úrovni. Systém je *o čosi* dôveryhodnejší než jeho časti.

Libertariánstvo sa potajomky spolieha na to, že väčšna jednotlivcov je dosť prosociálna, aby nechali sprepitné v reštaurácii, ktorú už nikdy nenavštívia. Ekonomika skutočne sebeckých činiteľov na ľudskej úrovni by sa zrútila. Podobne, Veda sa spolieha na to, že väčšina vedcov nebude páchať také mimoriadne hriechy, že si ich nebudú môcť racionalizovať.

Do tej miery, do akej vedci veria, že môžu propagovať svoje teórie pomocou akademického politizovania – alebo manipulovania štatistických metód, aby mali šancu vyhrať bez šance prehrať – alebo do tej miery, do akej sa nikto neunúva replikovať výsledky – efektivita vedy upadá. Avšak upadá elegantne, ako sa toto deje.

Tá časť, že úspešné predpovede patria tým teóriám a teoretickom, ktorí ich pôvodne urobili, a nemôže ich ukradnúť teória, ktorá príde neskôr – *bez* novej experimentálnej predpovede – je dôležitá súčasť tohto spoločenského procesu.

Konečným výsledkom je, že Veda sa nedá jednoducho zlúčiť s teóriou pravdepodobnosti. Ak urobíte pravdepodobnostný výpočet *správne*, dostanete *rozumnú* odpoveď. Veda nedôveruje vašej rozumnosti, a nespolieha sa na vašu schopnosť použiť teóriu pravdepodobnosti ako merítko pravdy. Vyžaduje od vás, aby ste zostavili definitívny experiment.

Považovať Vedu za púhu aproximáciu nejakého pravdepodobnostného ideálu rozumnosti... by iste vyzeralo *rozumne*. Zdá sa, že máme extrémne rozumne znejúci argument, že Bayesova veta je skrytá štruktúra, ktorá vysvetľuje, prečo Veda funguje. Lenže podriadiť Vedu veľkej schéme Bayesiánstva, a nechať Bayesiánstvo prísť a prebiť odpoveď Vedy, keď to vyzerá vhodné, nie je jednoduchý krok!

Veda je postavená na predpoklade, že aj vy ste *príliš hlúpy a sebaklamúci* na to, aby ste len tak použili Solomonoffovu indukciu. Napokon, keby to bolo také jednoduché, nepotrebovali by sme spoločenský proces vedy... však?

Takže, budete predsa len veriť vo víly kvantového „kolapsu“ rýchlejšieho než svetlo? Alebo si myslíte, že ste bystrejší?

*→

246. Keď veda nemôže pomôcť

Kde bolo, tam bolo, jeden mladý Eliezer mal jednu hlúpu teóriu. Povedzme, že hlúpa teória Eliezera$_{18}$ bola, že vedomie je spôsobené uzavretými časovými slučkami ukrytými v kvantovej gravitácii. Toto nie je celý príbeh, ani len približne, ale na začiatok nám to postačí.

Raz prišiel bod, keď som sa obzrel naspäť a uvedomil som si:

1. Počas svojho blúdenia som sa starostlivo riadil všetkým, o čom mi povedali, že je Tradične Rozumné. Napríklad som si dával pozor, aby som veril iba v také hlúpe teórie, ktoré robia nové experimentálne predpovede, napríklad že sa zistí, že mikrotubuly v neurónoch podporujú koherentné kvantové stavy.
2. Veda by bola dokonale spokojná s tým, keby som strávil desať rokov snahou otestovať svoju hlúpu teóriu, len aby som dostal negatívny výsledok, pokiaľ by som potom povedal: „Ach, jaj, moja teória bola asi nesprávna.“

Z pohľadu Vedy veci *majú* fungovať presne takto – zábava pre všetkých. Priznali ste si chybu! Ste úžasný! Nie je o tomto celá Veda?

Ale čo keby som nechcel premárniť desať rokov?

No... k *tomuto* mi Veda nemala veľmi čo povedať. Ako by Veda mohla povedať, ktorá teória je správna, *pred* urobením experimentálneho testu? Vedu nezaujíma, odkiaľ pochádza vaša teória – hovorí iba: „Choď to otestovať.“

Toto je veľká sila Vedy, a zároveň aj jej veľká slabosť.

Gray Area sa pýtal:

> Eliezer, prečo ťa zaujímajú netestovateľné otázky?

Pretože v otázkach, ktoré sa *ľahko ihneď* testujú, sa Veda ťažko pomýli.

Myslím tým, samozrejme, keď už máte k dispozícii jasné nezameniteľné experimentálne indície, riaďte sa nimi. Prečo by ste to nerobili?

Lenže niekedy bude mať otázka veľmi veľké, veľmi jasné experimentálne dôsledky vo vašej budúcnosti – ale *práve teraz* ju nedokážete jednoducho experimentálne otestovať – a predsa tu *je* silný *rozumný* argument.

→ http://lesswrong.com/lw/qb/science_doesnt_trust_your_rationality/

Makroskopické kvantové superpozície sú pripravené na testovanie: Potrebujete akurát nanotechnologickú presnosť, veľmi nízke teploty, a peknú prázdnu oblasť medzihviezdneho priestoru. No iste, nemôžete to urobiť *teraz hneď*, pretože je to *príliš drahé* alebo *nemožné s dnešnou technológiou* alebo niečo podobné... ale teoreticky, samozrejme! Ktovie, možno jedného dňa budú celé civilizácie bežať na kvantových počítačoch s makroskopickou superpozíciou, ďaleko v dôkladne upratanej oblasti Veľkej Prázdnoty. (Pýtať sa, čo hovorí kvantový nerealizmus o postavení pozorovateľov vnútri týchto počítačov, pomáha odhaliť nedostatočnú konkrétnosť kvantového nerealizmu.)

Toto nevyzerá ako okamžite pragmaticky relevantné pre váš život, tipol by som, ale ukazuje to vzor: Nie všetko, čo bude mať v budúcnosti dôsledky, sa dá *lacno* otestovať *teraz*.

Evolučná psychológia je ďalším príkladom situácie, kde rozumnosť musí prebrať rolu vedy. Hoci teórie evolučnej psychológie tvoria prepojený celok, iba niektoré z týchto teórií sme pripravení testovať experimentálne. Ale aj tak potrebujete tie zvyšné časti danej teórie, pretože tvoria navzájom prepojenú sieť, ktorá vám pomáha formulovať tie naozaj testovateľné hypotézy – a tie pomocné hypotézy sú teda podporené v bayesovskom zmysle, ale nie sú podporené experimentálne. Veda by mala vyniesť rozsudok „nedokázané“ nad jednotlivými časťami prepojenej teoretickej štruktúry, ktorá je experimentálne produktívna ako celok. Potrebovali by sme tu celkom nový druh rozsudku, niečo ako „nepriamo podporené“.

A čo kryonika?

Kryonika je archetypálnym príkladom extrémne dôležitej záležitosti (denne zomrie 150 000 ľudí), ktorá bude mať obrovské dôsledky v dohľadnej budúcnosti, ale neposkytuje jasné nezameniteľné experimentálne indície, ktoré by sme mohli mať *hneď teraz*.

Poviete teda: „Neverím v kryoniku, pretože nebola experimentálne dokázaná, a ľudia by nemali veriť na veci, ktoré neboli experimentálne dokázané?“

No, z bayesovského pohľadu je toto nesprávne. Neprítomnosť indície je indíciou neprítomnosti iba do tej miery, do akej by sme mohli rozumne očakávať, že sa indícia objaví. Ak niekto vyhlasuje, že hadí olej lieči rakovinu, môžete rozumne očakávať, že *keby hadí olej naozaj liečil rakovinu*, nejaký vedec by robil kontrolovanú štúdiu, aby to overil – že by prinajmenšom lekári ohlasovali prípadové štúdie zázračných uzdravení – takže neprítomnosť takejto indície je silnou indíciou neprítomnosti. Ale „medzery v zázname skamenelín“ nie sú silnou indíciou proti evolúcii; skameneliny vznikajú iba vzácne, a *dokonca aj keby prechodný vývojový článok naozaj existoval*, nemôžete očakávať s veľkou pravdepodobnosťou, že z neho Príroda poslušne vyrobí skamenelinu, a že túto skamenelinu objavíme.

Oživiť kryonicky zamrazeného cicavca nie je niečo, čo by ste očakávali, že sa bude dať pomocou modernej technológie, *dokonca ani keby budúce nanotechnológie naozaj dokázali vykonať úspešné oživenie.* Takto to podľa mňa vychádza podľa Bayesa.

Aha, a čo sa táka skutočných argumentov *za* kryoniku – teraz sa do nich nejdem púšťať. Ale ak ste sledovali postupnosti o fyzike a proti zombiám, malo by teraz vyzerať o dosť dôveryhodnejšie, že to, čo zachová vzor synapsií, zachová z „vás“ rovnako veľa, ako sa zachová od večerného zaspania do ranného prebudenia.

V rámci férovosti, keď niekto povie: „Neverím v kryoniku, pretože nebola experimentálne dokázaná“, *nesprávne aplikuje* pravidlá Vedy; toto je nie je prípad, kde veda naozaj dáva *nesprávnu odpoveď*. V neprítomnosti jasného experimentálneho testu je rozsudok vedy „nedokázané“. Ak to niekto interpretuje ako odmietnutie, to je jeho krok navyše mimo vedy, nie prešľap v rámci vedy.

Stránka wikicitátov cituje Johna McCarthyho: „Vaše tvrdenia sú ako povedať, že ak je UI možná, mala by byť jednoduchá. Prečo?“[210] Stránka wikicitátov neuvádza, na čo McCarthy reagoval, ale môžem sa pokúsiť hádať.

Táto všeobecná chyba pravdepodobne vzniká, pretože *existujú* prípady, kde neprítomnosť vedeckého dôkazu je silnou indíciou – pretože experiment by sa dal ľahko zostaviť, takže jeho nezostavenie je samo osebe podozrivé. (Hoci nie až také podozrivé, ako som si kedysi myslel – keď z

210 Už to nie je na wikicitátoch, ale je to na McCarthyho stránke citátov.

uznávaných zdrojov prichádzajú všetky tie zvláštne rozmanité anekdotálne indície, prečo *dočerta* nikto netestuje teórie Setha Robertsa o potláčaní chuti?[211])

Ďalší mätúci činiteľ môže byť, že keď testujete Liek X na 1000 pokusných osobách a zistíte, že sa uzdraví 56 % kontrolnej skupiny a 57 % experimentálnej skupiny, niektorí ľudia to označia za rozsudok „nedokázané“. Ja by som to označil za experimentálny rozsudok „Liek X nefunguje dobre, ak vôbec“. To, že tento rozsudok teoreticky možno odvolať v prípade nových indícií, ešte neznamená, že je nejednoznačný.

Každopádne tu teraz máme ľudí, ktorí spakruky odmietajú kryoniku ako „nevedeckú“, ako keby to bol nejaký druh lieku, ktorý môžete ľahko podať 1000 pacientom a pozrieť, čo sa stane. „Zavolajte mi, až kryonika naozaj niekoho oživí“, hovoria; o čom Mike Li konštatuje, že to je ako povedať: „Odmietam nasadnúť do tejto sanitky; zavolajte mi, až bude naozaj v nemocnici.“ Možno ich Martin Gardner varoval, aby neverili čudným veciam bez experimentálnej indície. Čakajú teda na definitívny nezameniteľný rozsudok Vedy, zatiaľ čo ich rodina a priatelia a 150 000 ľudí denne zomierajú *práve teraz*, a možno by sa dali a možno nedali zachrániť...

...čo je stávka, ktorú môžete vypočítať iba *rozumom*.

Motorom Vedy je zháňanie takých obrovských hôr indície, že ich dokonca ani omylní ľudskí vedci nedokážu prečítať nesprávne. Ale aj *to* sa občas pokazí, keď si ľudia popletú, ktorá teória čo predpovedá, alebo zakomponujú do raných verzií svojej teórie extrémne ťažko testovateľné časti. A niekedy skrátka nedostanete vôbec žiadne jasné experimentálne indície.

Tak či onak, musíte sa pokúsiť urobiť to, v čom Veda nikomu nedôveruje – myslieť rozumne, a zistiť si odpoveď skôr než vám bude oplieskaná o hlavu.

(Ach, a občas *vyvracajúci* experimentálny výsledok vyzerá takto: „Celý váš živočíšny druh bol práve vyhubený! Teraz ste vedecky povinní vzdať sa svojej teórie. Ak ju verejne odvoláte, ste super! Pamätajte, že iba silná myseľ sa dokáže vzdať dávno držaných názorov. Nech sa páči, nabudúce môžete skúsiť inú hypotézu!“)

* →

247. Veda nie je dosť prísna

Kde bolo, tam bolo, jeden mladý Eliezer mal jednu hlúpu teóriu. Eliezer$_{18}$ sa opatrne riadil predpismi Tradičnej Rozumnosti, ktoré ho učili; dal si pozor na to, aby jeho hlúpa teória mala experimentálne dôsledky. Eliezer$_{18}$ vyznával, v súlade s vedeckými cnosťami, ktoré ho naučili, že si želá otestovať svoju hlúpu teóriu.

To bolo všetko, čo sa od neho v mene cnosti vyžadovalo, podľa toho, čo Eliezera$_{18}$ učili ohľadom vedeckých cností.

Nebolo to ani *zďaleka* toľké úsilie, koľko by bolo treba, aby to robil *správne*.

Tradičné myšlienky Vedy príliš ľahko rozdávajú zlaté hviezdičky. Negatívne experimentálne výsledky sú tiež poznanie, takže každý, kto sa zapojí do hry, získava cenu. Pokiaľ si dokážete vymyslieť nejaký experiment, ktorý testuje vašu teóriu, a *urobíte* tento experiment, a *prijmete* jeho výsledky, hrali ste podľa pravidiel; ste dobrý vedec.

Nemuseli ste to nutne mať správne, ale ste vzorný, vedu poslúchajúci občan.

(V tomto bode poznamenávam, že hovorím o Vede, nie o spoločenskom procese vedy ako naozaj funguje v praxi, z dvoch dôvodov. Po prvé, zablúdil som, keď som sa snažil riadiť *ideálmi* Vedy – nešlo o to, že by ma odstrelil nejaký redaktor časopisu, korému som sa nepáčil, alebo že by som sa snažil

211 Seth Roberts, „What Makes Food Fattening?: A Pavlovian Theory of Weight Control“ [Prečo sa z jedla priberá?: Pavlovovská teória kontroly hmotnosti] (Unpublished manuscript, 2005), http://media.sethroberts.net/about/-whatmakesfoodfattening.pdf.

→ http://lesswrong.com/lw/qc/when_science_cant_help/

napodobňovať chyby akadémie. Po druhé, ak ukážem na problém v tradične vyučovanom ideále, neznamená to, že vedci v skutočnom živote by boli *nútení* blúdiť rovnakým spôsobom!)

Veda začala ako vzbura proti veľkým filozofickým konštrukciám a jalovému špekulovaniu. Preto Veda neobsahuje pravidlo o tom, aké druhy hypotéz smiete alebo nesmiete testovať; to záleží od rozhodnutia jednotlivého vedca. Pokúsiť sa to uhádnuť a priori, to by si vyžadovalo nejakú veľkú filozofickú konštrukciu a špekulovanie pred získaním indícií. Ako spoločenský ideál vás Veda neodsúdi ako zlého človeka za navrhnutie kacírskej hypotézy; poctivé experimenty a zmierenie sa s ich výsledkami, to je cnosťou vedca.

Dokiaľ sa väčšina vedcov dokáže zmieriť s jasnými, nezameniteľnými, jednoznačnými experimentálnymi indíciami, veda sa môže rozvíjať. Môže sa to diať pomaly – môže to trvať dlhšie než by malo – možno musíte počkať, kým vymrie staršia generácia – ale nakoniec západka poznania poskočí dopredu o jeden zub. Rok za rokom, desaťročie za desaťročím sa koleso otáča *vpred*. To stačí na udržanie civilizácie.

Takže toto je všetko, čo od vás Veda v skutočnosti požaduje – schopnosť zmieriť sa so skutočnosťou, keď vám ju oplieskajú o hlavu. Nie je to veľa, ale stačí to na udržanie vedeckej kultúry.

Porovnajte si to s predstavou, ktorú máme v teórii pravdepodobnosti, kvantitatívne presného rozumného úsudku. Ak 1 % žien, ktoré prídu na rutinné vyšetrenie, má rakovinu prsníka, a 80 % žien s rakovinou prsníka dostane pozitívny mamogram, a 10 % žien bez rakoviny prsníka dostane falošný potizívny výsledok, aká je pravdepodobnosť, že rutinne vyšetrená žena s pozitívnym mamogramom má rakovinu prsníka? 7,5 %. Nemôžete povedať: „Myslím si, že nemá rakovinu prsníka, lebo ten experiment nebol dosť dôkladný.“ Nemôžete povedať: „Myslím si, že má rakovinu prsníka, pretože je múdre byť pesimistom, a toto je to, čo naznačuje zatiaľ jediný experiment.“ Rozumný odhad pri tejto indícii je 7,5 %, nie 7,4 %, ani 7,6 %. Zákony pravdepodobnosti sú *zákony*.

V *Dvanástich cnostiach* sa píše o tretej cnosti, ľahkosti:

> Ak vnímaš indície ako obmedzenia a usiluješ sa oslobodiť, predávaš sám seba do reťazí svojich rozmarov. Nemôžeš totiž nakresliť pravdivú mapu mesta tak, že sedíš vo svojej spálni so zatvorenými očami a kreslíš čiary na papier podľa popudov. Musíš prejsť cez mesto a kresliť na papier čiary zodpovedajúce tomu, čo vidíš. Ak vidíš mesto nejasne a pomyslíš si, že môžeš posunúť nejakú čiaru len o kúsok doprava, len o kúsok doľava, podľa svojho rozmaru, toto je rovnaká chyba.

Keď vo Vede príde na rozhodovanie, ktoré hypotézy testovať, morálka Vedy vám dáva osobnú slobodu, čomu chcete veriť, dokiaľ to už nie je vylúčené experimentom, a dokiaľ prejdete k testovaniu svojej hypotézy. Veda by sa nepokúsila dať oficiálny rozsudok, ktorú hypotézu je *najlepšie* testovať, ešte *pred* experimentom. To je ponechané na svedomie jednotlivého vedca.

Kde existujú definitívne experimentálne indície, Veda vám povie, aby ste sklonili svoju tvrdú hlavu a zmierili sa s tým. Inak to Veda necháva na vás. Veda vám dáva priestor potulovať sa po okolí *v rámci hraníc* experimentálnej indície podľa vášho rozmaru.

A nie je ľahké zmieriť toto s predstavou Bayesiánstva o presne správnom odhade pravdepodobnosti, v ktorom nie je miesto na rozmary, ktorý existuje aj pred experimentom, aj po ňom. Nezapadá to do prastarého a tradičného dôvodu pre Vedu – nedôvera vo veľké schémy, predpoklad, že ľudia nie sú dosť rozumní na to, aby došli na veci správne bez definitívnej a nezameniteľnej experimentálnej indície. Keby sme všetci boli dokonalými Bayesovcami, *nepotrebovali* by sme spoločenský proces vedy.

Napriek tomu, zhruba v čase, keď som si uvedomil svoju veľkú chybu, čítal som aj Kahnemana a Tverskyho a Jaynesa. Učil som sa novú Cestu, prísnejšiu než je Veda. Cestu, ktorá by kritizovala moje bláznovstvo spôsobom, ktorým by Veda nemohla. Cestu, ktorá by mi povedala to, čo by Veda nikdy nepovedala *vopred*: „Vybral si si nesprávnu hypotézu na testovanie, trdlo.“

Lenže Bayesova Cesta sa aj *používa omnoho ťažšie* než Veda. Ukladá ťažké bremeno na vašu schopnosť počuť drobné falošné tóny tam, kde Veda iba žiada, aby ste si všimli, keď vám padne nákova na hlavu.

Vo Vede môžete urobiť jednu alebo dve chyby, a príde ďalší experiment a opraví vás; v najhoršom prípade premárnite niekoľko desaťročí.

Ale ak sa pokúsite používať Bayesa hoci len kvalitatívne – ak sa pokúsite urobiť to, čo Veda nedôveruje, že dokážete, čiže rozmýšľať rozumne v neprítomnosti drvivej indície – je to ako matematika, v tom, že jediná chyba v sto krokoch vás môže doviesť k hocičomu. Vyžaduje si ľahkosť, vyrovnanosť, presnosť, perfekcionizmus.

Existuje dobrý dôvod, prečo Veda nedôveruje vedcom v robení takýchto vecí, a žiada si ďalšie experimentálne dôkazy *aj potom*, čo niekto tvrdí, že našiel správnu odpoveď na základe náznakov a logiky.

Ale ak by ste radšej nestrácali desať rokov snahou dokázať *nesprávnu* teóriu, musíte sa podujať na omnoho náročnejšiu úlohu: počúvať indície, ktoré vám nekričia do ucha.

Aj keď si nedokážete nájsť pôvodné pravdepodobnosti pre nejaký problém v *Chemických a fyzikálnych tabuľkách* – dokonca aj keď vám žiaden Autoritatívny Zdroj nehovorí, aké sú tieto pôvodné údaje – to neznamená, že si môžete slobodne, osobne vybrať akékoľvek pôvodné údaje chcete. Znamená to, že máte novú hádanku, ktorú musíte vyriešiť podľa svojich najlepších schopností.

Keby myseľ, ako stroj na poznanie, dokázala vytvárať *správne* odhady tým, že by sa pohrávala s pôvodnými údajmi podľa svojich rozmarov, mohli by ste vedieť veci skôr než by ste sa na ne pozreli, alebo ich dokonca meniť bez toho, aby ste sa ich dotkli. Lenže myseľ nie je čarovná. V rozumnom odhade pravdepodobnosti nie je žiaden priestor pre rozmar, dokonca ani keď sa vám zdá, že neviete pôvodné pravdepodobnosti.

Podobne, ak je bayesovskú odpoveď ťažké vypočítať, neznamená to, že nemožno použiť Bayesa; znamená to, že *neviete*, aká je bayesovská odpoveď. Bayesovská teória pravdepodobnosti nie je zbierka štatistických metód, je to *zákon*, ktorým sa riadia všetky nástroje, ktoré používate, či o tom viete alebo neviete, či to počítate alebo nie.

Čo sa týka používania bayesovských metód na obrovské, veľmi všeobecné priestory hypotéz – ako napríklad: „Tú sú údaje z celkom všetkých fyzikálnych experimentov; aká by teda bola dobrá Teória Všetkého?“ - keby ste vedeli toto urobiť *v praxi*, neboli by ste štatistik, boli by ste programátor Všeobecnej Umelej Inteligencie. Ale to neznamená, že ak ľudia modelujú vesmír pomocou ľudskej inteligencie, že tým porušujú fyzikálne zákony / Bayesiánstvo vytváraním dobrých odhadov bez indícií.)

Nick Tarleton povedal:

> Problémom je podporovanie rovnako nedbanlivého *súkromného* štandardu *poznania* ako je ten spoločenský.

čo vystihuje problém, na ktorý som sa tu snažil ukázať, omnoho lepšie než ja.

* →

248. Vedia už vedci o tomto?

poke tvrdí:

> „Dokázať vytvoriť relevantnú hypotézu je dôležitá zručnosť, ktorej rozvíjaním strávi vedec veľa zo svojho času. Možno to nie je súčasť tradičného *popisu* vedy, ale to neznamená, že to nie je súčasťou skutočnej spoločenskej inštitúcie vedy, ktorá vytvára skutočnú vedu v našom skutočnom svete; chyba je v tvojom popise a nie vo vede.“

Viem, že som označil svoje ja ako „hlúpe“, ale to mal byť rečnícky prostriedok; presnejšie by bolo: „nešikovne zaobchádzajúce s vysokou inteligenciou“. Eliezer[18] nemal vo zvyku robiť samozrejmé chyby – akurát jeho „samozrejmé“ nebolo rovnaké ako moje „samozrejmé“.

→ http://lesswrong.com/lw/qd/science_isnt_strict_enough/

Nie, neprešiel som tradičným učňovstvom. Ale keď sa ozbriem a vidím, čo robil Eliezer$_{18}$ nesprávne, vidím, že *mnohí* moderní vedci robia tie isté chyby. Neviem nájsť žiadne znamenie, že by boli varovaní lepšie než ja.

Sir Roger Penrose – svetový fyzik – si stále myslí, že vedomie spôsobuje kvantová gravitácia. Myslím si, že ho nikto nikdy nevaroval pred tajomnými odpoveďami na tajomné otázky – iba mu povedali, že jeho hypotéza musí byť falzifikovateľná a musí mať empirické dôsledky. Tak ako Eliezerovi$_{18}$.

„Vedomie spôsobuje kvantová gravitácia“ má testovateľné dôsledky: Vyplýva z toho, že by ste sa mali dokázať pozrieť na neurón a objaviť koherentnú kvantovú superpozíciu, ktorej kolaps prispieva k spracovaniu informácií, a že by sa vám nikdy nemalo podariť reprodukovať vstupno-výstupné správanie neurónu pomocou spočítateľnej mikroanatomickej simulácie...

...ale aj potom, čo poviete: „Vedomie spôsobuje kvantová gravitácia“, neočakávate nič ohľadom toho, ako si váš mozog myslí: „Myslím, teda som!“ alebo ohľadom tajomnej červenosti červenej, čo by ste neboli očakávali aj predtým, hoci teraz máte pocit, že už poznáte príčinu za tým. Toto je ohromné varovné znamenie, ako si *uvedomujem teraz,* ale nie je to varovné znamenie, na ktoré by *mňa* niekto upozornil, a pochybujem, že to niekedy Penroseovi povedal jeho školiteľ. Keď už sme pri tom, pochybujem, že pred tým niekedy varovali Nielsa Bohra, keď prišiel čas formulovať Copenhagenskú interpretáciu.

Pokiaľ viem povedať, Eliezer$_{18}$ ani Sir Roger Penrose ani Niels Bohr neboli varovaní preto, lebo žiadne štandardné varovanie neexistuje.

Nezovšeobecnil som si pojem „tajomných odpovedí na tajomné otázky“ toľkými slovami, dokiaľ som nepísal bayesovskú analýzu o tom, čím sa líšia technické, netechnické a polotechnické vedecké vysvetlenia. *Výsledok* tejto analýzy sa teraz dá vyjadriť netechnicky pomocou štyroch varovných znamení:

- Po prvé, vysvetlenie funguje ako zastavovač zvedavosti a nie ako ovládač očakávania.
- Po druhé, hypotéza nemá žiadne pohyblivé časti – tajomnou prísadou nie je konkrétny zložitý mechanizmus, ale prázdna rovnomerná podstata alebo sila.
- Po tretie, tí, ktorí ponúkajú toto vysvetlenie, milujú svoju nevedomosť; rozprávajú hrdo o tom, ako tento jav odporuje bežnej vede alebo je celkom iný ako púhe svetské javy.
- Po štvrté, *ešte aj po tom, čo dostanete odpoveď, je tento jav stále tajomný* a obsahuje rovnakú vlastnosť úžasnej nevysvetliteľnosti, ako mal na začiatku.

V princípe by sa toto všetko dalo povedať okamžite po víťazstve vitalizmu. Rovnako ako by základnú teóriu pravdepodobnosti mohol objaviť Archimedes, alebo ako by starovekí Gréci mohli vyvinúť teóriu prirodzeného výberu. Lenže *v skutočnosti* ma nikto nevaroval pred žiadnym z týchto štyroch nebezpečenstiev, týmito slovami – najbližšie k tomu bolo varovanie, že hypotézy by mali mať testovateľné dôsledky. A nepreložil som si tieto varovné znamenia *do slov*, dokiaľ som sa nepokúsil myslieť na celú túto záležitosť v pojmoch distribúcie pravdepodobnosti – bolo treba zájsť omnoho ďalej.

Jednoducho nemám dôvod veriť, že sa tieto varovania odovzdávajú pri vedeckom učňovstve – rozhodne nie u väčšiny vedcov. Okrem iných vecí je to rada na *zvládanie situácií zmätku a zúfalstva*, vedeckého *chaosu*. Kedy má priemerný vedec alebo priemerný školiteľ príležitosť použiť takýto druh techniky?

Práve sme sa dostali na záver diskusie o fiasku jedného sveta vo fyzike. Je zrejmé, že im nikto nepovedal o formálnej definícii Occamovej britvy, pošepky v rámci učňovstva alebo inak.

Je známy efekt, že špičkoví vedci majú mnohých špičkových študentov. To ľahko môže byť vďaka tomu, že školitelia odovzdávajú zručnosti, ktoré nevedia opísať. Ale nemyslím si, že sa toto počíta ako časť *štandardnej* vedy. A ak títo špičkoví školitelia nedokázali vložiť svoje vedenie do slov a uverejniť ho všeobecne, to nie je dobré znamenie ohľadom kvality pochopenia týchto vecí.

Uvažovanie v neprítomnosti definitívnych indícií bez *okamžitého úplného mýlenia sa* je *naozaj naozaj ťažké*. Keď sa učíte v škole, môžete si nevšimnúť jednu vec, a potom sa naučiť päťdesiat ďalších

vecí, ktoré budú správne. Keď rozumom vytvárate nové vedomosti v neprítomnosti drvivo valcujúcich síl, môžete sa pomýliť v jednej veci a prebudiť sa o päťdesiat krokov neskôr v Mongolsku.

Som si celkom istý, že vedci, ktorí vypínajú svoje mozgy a relaxujú s nejakým príjemným nezmyslom akonáhle opustia svoju špecializáciu, si neuvedomujú, že mozgy sú stroje a že za každým dôveryhodným názorom je kauzálny príbeh. Mám podozrenie, že im ani nikdy nepovedali, že pri danom stave indície existuje presná rozumná pravdepodobnosť, v ktorej nie je miesto na rozmary; dokonca aj keď nedokážete spočítať túto odpoveď, a dokonca aj keď nepočujete žiaden príkaz autority, čomu máte veriť.

Pochybujem, že vedcov, ktorých médiá žiadajú o pontifikát ohľadom budúcnosti, ktorí vykresľujú úžasne podrobné obrázky Života v roku 2050, niekedy učili o klame konjunkcie. Alebo ako heuristika reprezentatívnosti môže urobiť podrobnejší príbeh uveriteľnejším, hoci každá podrobnosť navyše znižuje jeho pravdepodobnosť. Samotná predstava, že každá pridaná podrobnosť potrebuje svoju vlastnú podporu – že nemôžeme *vymýšľať* veľké príbehy plné podrobností, ktoré znejú rovnako ako tie príbehy plné podrobností, ktoré vás *učili* na hodinách fyziky alebo dejepisu – je *absolútne životne dôležitá* na presné myslenie v neprítomnosti definitívnych indícií. Ale ako by sa takáto predstava dostala do *štandardného* vedeckého učňovstva? Toto kognitívne skreslenie bolo objavené iba pred pár desaťročiami, a až donedávna nebolo popularizované.

Ďalej existujú afektívne špirály smrti okolo pojmov ako „emergencia“ alebo „zložitosť“, ktoré sú dostatočne hmlisto definované, aby ste o nich mohli povedať veľa pekných vecí. Okolo toho druhu chýb, aké robil Eliezer$_{18}$, sú postavené celé akademické podoblasti! (Hoci ja som sa nikdy nedal nachytať na „emergenciu“.)

Niekedy hovorím, že cieľom vedy je zhromaždiť také ohromné hory indície, že ich nedokážu ignorovať ani len vedci: a že toto je charakteristickou črtou vedca, pretože nevedec to aj tak bude ignorovať.

Ak môže existovať nejaké množstvo indície také zdrvujúce, že nakoniec v zúfalstve prestanete robiť výhovorky a *jednoducho sa vzdáte* – zahodíte starú teóriu a nikdy viac ju nespomeniete – toto je všetko, čo treba, aby sa západka Vedy časom otáčala dopredu, a vybudovala technologickú civilizáciu. Na rozdiel od náboženstva.

Knihy od Carla Sagana a Martina Gardnera a ďalších postáv Tradičnej Rozumnosti mali za úlohu dosiahnuť tento rozdiel: premeniť niekoho z nie-vedca na potenciálneho vedca, a chrániť ho pred experimentálne vyvráteným šialenstvom.

Aký ďalší výcvik dostane profesionálny vedec? Nejaké lekcie frekventistickej štatistiky o tom, ako sa počíta štatistická významnosť. Výcvik v štandardných technikách im umožní chrliť články bez pevne podloženej paradigmy.

Keby Veda žiadala viac než toto od priemerného vedca, nemyslím si, že by bolo možné robiť Vedu. Máme dosť problémov s ľuďmi, ktorí sa prešmyknú dnu bez základnej kvalifikácie na prežitie.

Nick Tarleton zhrnul výsledný problém veľmi dobre – vlastne lepšie než ja: Ak prídete s čudesne vyzerajúcou hypotézou, ktorú zatiaľ indície nevyvrátili, a pokúsite sa ju experimentálne testovať, Veda vás neoznačí za zlého človeka. Veda nedôveruje svojim starešinom pri rozhodovaní, ktoré hypotézy „sa neoplatí testovať“. Ale toto je úmyselne nedbanlivý *spoločenský* štandard, a ak sa z neho pokúsite urobiť štandard epistemickej rozumnosti *jednotlivca*, dovolí vám veriť priveľa. Aby som sa vrátil k analógii s libertariánstvom založeným na pragmatickej nedôvere, je to ako rozdiel medzi „Cigarety by nemali byť nelegálne“ a „Choď si zapáliť Marlboro“.

Pamätáte sa, že by vás niekedy niekto *varoval pred tou chybou*, toľkými slovami? Prečo by potom ľudia *nerobili* presne tú chybu? Koľkí ľudia *spontánne* prejdú míľu navyše, a budú voči sebe ešte prísnejší? Niektorí, ale nie mnohí.

Mnohí vedci budú mimo laboratória veriť všemožným smiešnym veciam, dokiaľ dokážu presvedčiť sami seba, že to nebolo definitívne vyvrátené, alebo dokiaľ si dokážu nepoložiť túto otázku. Existuje

nejaká štandardná lekcia pre doktorandov, podľa ktorej ľudia vidia túto pochabosť a pýtajú sa: „Vari si v ten deň chýbal na hodine?“ Pokiaľ viem, nie.

Možno ak máte super šťastie a dostanete slávneho školiteľa, povedia vám vzácne osobné tajomstvá ako: „Polož si otázku, ktoré sú najdôležitejšie problémy v tvojom odbore, a potom pracuj na jednom z nich, namiesto zakopania sa v niečom jednoduchom a triviálnom“ alebo „Dávaj si väčší pozor než vyžadujú redaktori odborných časopisov; hľadaj nové spôsoby, ako chrániť experimenty pred vplyvom tvojich očakávaní, aj tam, kde sa to štandardne nerobí.“

Ale *naozaj si nemyslím*, že existuje veľká tajná štandardná vedecká tradícia presného rozumného uvažovania pri nedostatku indície. Polovica všetkých vedcov na svete stále verí, že verí v Boha! *Náročnejšie zručnosti nie sú štandard!*

*→

249. Žiadne bezpečné útočisko, ani len veda

Nezvyknem sa pýtať svojich priateľov na ich detstvo – chýba mi spoločenská zvedavosť – a tak neviem, nakoľko je toto naozaj bežné:

Medzi mojimi známymi, ktorí sa snažia rásť ako racionalisti, a ktorí poskytnú informácie o svojom detstve, je prekvapivo často počuť veci ako: „Moja rodina vstúpila do sekty a ja som z nej musel vystúpiť“ alebo „Jeden z mojich rodičov bol duševne chorý a ja som sa musel naučiť filtrovať skutočnosť od jeho šialenstva.“

Moja osobná skúsenosť s vyrastaním v ortodoxnej židovskej rodine je v porovnaní s tým neškodná... ale dosiahla ten istý výsledok: Zlomila moju základnú emocionálnu dôveru v príčetnosť ľudí okolo mňa.

Dokiaľ sa nezlomí táto emocionálne dôvery, nezačnete rásť ako racionalista. Mám problém vyjadriť slovami, prečo je to tak. Možno ľubovoľná *nezvyčajná* zručnosť, ktorú získate – čokoľvek, čo vás urobí *nezvyčajne* rozumným – si vyžaduje, aby ste urobili cik, keď ostatní robia cak. Možno je to len príliš desivé, ak vám svet stále pripadá ako príčetné miesto.

Alebo sa možno neunúvate makať na tom, aby ste boli extra navyše príčetní, pokiaľ vás normálnosť nedesí do špiku kostí.

Viem, že mnoho ašpirujúcich racionalistov napohľad narazilo na bariéru okolo vecí ako kryonika alebo mnohé svety. Nie, že by nevideli tú logiku; vidia tú logiku a čudujú sa: „Môže to naozaj byť pravda, keď sa mi to teraz zdá také samozrejmé, a predsa tomu neverí nikto okolo mňa?“

Áno. Vitajte na Zemi, kde sa etanol vyrába z kukurice a ochrancovia životného prostredia protestujú proti jadrovej energii. Mrzí ma to.

(Pozrite si aj: Sektárske antisektárstvo. Ak skončíte v stave mysle, že *nervózne hľadáte útechu*, to nikdy nie je dobrá vec – dokonca aj keď je to preto, že sa chystáte uveriť niečomu, čo znie logicky, ale mohlo by spôsobiť, že sa na vás druhí ľudia budú čudne pozerať.)

Ľudia, ktorých dôvera v príčetnosť ľudí okolo nich bola zlomená, vyzerajú byť schopnejší vyhodnocovať čudné myšlienky podľa ich zásluh, bez nervozity z ich čudnosti. Lepidlo, ktoré ich púta k ich súčasnému miestu sa rozpustilo, a oni môžu kráčať ľubovoľným smerom, v lepšom prípade dopredu.

Nazýval som to osamelý nesúhlas. Skutočný nesúhlas, to nie je ako chodiť do školy v čiernom; je to ako chodiť do školy v šašovskom obleku.

Tak sa cíti ten osamelý hlas, ktorý hovorí: „Ak naozaj vieš, kto vyhrá tieto voľby, prečo si nevezmeš peniaze zadarmo z burzy predpovedí Intrade?“ zatiaľ čo všetci ľudia okolo vás si myslia: „Je dobré byť jednotlivcom a vytvárať si svoje vlastné názory, lebo mi to povedala reklama na topánky.“

Možno v nejakom inom svete, v nejakej alternatívnej Everettovej vetve s príčetnejšou populáciou ľudí, by veci boli inak... ale v tomto svete som nikdy nevidel nikoho začať rásť ako racionalista, dokiaľ nezlomil hlboké puto s múdrosťou svojej svorky.

→ http://lesswrong.com/lw/qe/do_scientists_already_know_this_stuff/

Možno v inom svete by veci boli inak. A možno nie. Nie som si istý, či ľudia *dokážu* naozaj dôverovať aj rozmýšľať zároveň.

Kde bolo, tam bolo, bolo raz niečo, čomu som veril.

Eliezer$_{18}$ veril Vede.

Eliezer$_{18}$ poslušne priznával, že spoločenský proces vedy mal svoje chyby. Eliezer$_{18}$ poslušne priznával, že akadémia je pomalá, že zle hospodári, že sa hrá na protekcie, a že sa správa zle voči svojim vzácnym kacírom.

To je totiž to pohodlné na priznávaní chýb v *ľuďoch*, ktorí nežijú v súlade s vaším ideálom; nemusíte spochybňovať samotný ideál.

Ale kto už by mohol byť taký blázon, aby spochybňoval: „Experimentálna metóda rozhodne, ktorá hypotéza vyhrá?“

Časť toho, čo poplietlo Eliezera$_{18}$, bol jeho všeobecný problém, odpor voči veciam, ktoré mu pripomínali veci, ktoré povedali idioti. Eliezer$_{18}$ videl mnohých ľudí, ako spochybňujú ideály samotnej vedy, a tí boli všetci bez výnimky na Temnej Strane. Ľudia, ktorí spochybňovali ideál Vedy sa bez výnimky snažili predať vám hadí olej, alebo sa pokúšali ochrániť svoju obľúbenú formu hlúposti pred kritikou, alebo sa snažili zamaskovať svoje osobné rezignovanie ako Hlbokú Múdrosť zmierenia sa s márnosťou.

Keby existoval nejaký iný ideál, ktorý by mal pár storočí, mladý Eliezer by sa naň bol pozrel a povedal by: „Ktovie, či je toto naozaj pravda, a či existuje spôsob, ako to robiť lepšie.“ Nie však ideál Vedy. Veda bola tá najvyššia myšlienka, myšlienka, ktorá vám umožňuje meniť myšlienky. Mohli ste o nej pochybovať, ale očakávalo sa, že o nej zapochybujete a potom sa s ňou zmierite, nie že naozaj poviete: „Moment! Toto je zle!“

Keď som teda jedného dňa prišiel s hlúpou myšlienkou, myslel som si, že sa správam cnostne, ak si dám pozor, aby existovala Nová Predpoveď, a ak vyznám, že chcem, aby sa moja myšlienku experimentálne testovala. Myslel som si, že som urobil všetko, čo som bol povinný urobiť.

Myslel som si teda, že som v *bezpečí* – nie v bezpečí pred nejakou konkrétnou vonkajšou hrozbou, ale v bezpečí na akejsi hlbšej úrovni, ako dieťa, ktoré dôveruje svojim rodičom, a ktoré poslúchlo všetky rodičovské pravidlá.

Už dávno predtým sa mi zlomila dôvera v príčetnosť mojej rodiny alebo mojich školských učiteľov. A ostatné deti neboli dosť inteligentné na to, aby konkurovali rozhovorom, ktoré som mohol mať s knihami. Ale knihám som dôveroval, ako vidíte. Dôveroval som, že keď urobím to, čo mi Richard Feynman povedal, aby som urobil, že budem v bezpečí. Nikdy som si tie slová nepomyslel nahlas, ale tak som sa cítil.

Keď si Eliezer$_{23}$ uvedomil, *ako* presne hlúpa bola tá hlúpa teória – a že ho pred ňou Tradičná Rozumnosť neochránila – a že Vede by absolútne nevadilo, keby premárnil desať rokov testovaním tej hlúpej myšlienky, pokiaľ by nakoniec priznal, že sa mýlil...

...no, nechystám sa povedať, že to bol veľký emocionálny kŕč. Ja naozaj nemám sklon k tomuto typu drámy. Jednoducho mi prišlo zrejmé, že som bol hlúpy.

To je tá dôvera, ktorú skúšam zlomiť u vás. Nie ste v bezpečí. Nikdy.

Ani len Veda vás neochráni. Ideály Vedy vznikli pred stáročiami, v čase keď nikto nič nevedel o teórii pravdepodobnosti alebo o kognitívnych skresleniach. Veda od vás vyžaduje *príliš málo*, požehnáva vaše dobré úmysly príliš ľahko, nie je *dosť* prísna, robí iba také pravidlá, ktorými sa dokáže riadiť priemerný vedec, akceptuje pomalosť ako súčasť života.

Nemyslím si teda, že ak sa iba riadite pravidlami Vedy, že tým vaše uvažovanie možno obhájiť.

Nie je známa taká procedúra, ktorú by ste mohli nasledovať, aby sa vaše uvažovanie dalo obhájiť.

Nie je známa taká sada pravidiel, ktoré by ste mohli dodržať a potom by ste vedeli, že nebudete bláznom.

Nie je známa taká morálka uvažovania, ktorú by ste mohli zo všetkých síl poslúchať, a vedieť, že vás to ochráni pred kritikou.

Nie, dokonca ani keď sa dáte na bayesovské remeslo. Používa sa omnoho ťažšie a nikdy si nebudete istí, že to robíte správne.

Učenie Bayesovho remesla je omnoho mladšie než učenie Vedy. Nenájdete žiadne učebnice, žiadnych starších školiteľov, žiadne písané dejiny úspechov a zlyhaní, žiadne rýchle a jasné napísané pravidlá. Budete musieť študovať kongitívne skreslenia a teóriu pravdepodobnosti a evolučnú psychológiu a sociálnu psychológiu a iné kognitívne vedy a umelú inteligenciu – a myslieť sami za seba, ako uplatiť všetko toto poznanie na opravenie seba samého, pretože to ešte v tých učebniciach nie je.

Neviete, čo vaša vlastná myseľ naozaj robí. Každý týždeň objavia nejaké nové kognitívne skreslenie a nikdy si nie ste istí, či ste ho opravili, alebo ste to s opravou prehnali.

Formálna matematika sa tu nedá aplikovať. Nezlyháva až tak ľahko, ako si myslí John Q. Neveriaci, ale nikdy si nie ste naozaj istí, odkiaľ pochádzajú základy. Neviete, prečo je vesmír dosť jednoduchý na to, aby sa dal pochopiť, alebo prečo v ňom fungujú vôbec nejaké pôvodné predpoklady. Neviete, aké *sú* vaše pôvodné predpoklady, ani či sú vôbec na niečo dobré.

Jeden z problémov Vedy je, že je príliš nejasná na to, aby vás naozaj vydesila. „Myšlienky treba testovať experimentom.“ Čo sa na tomto dá pokaziť?

Na druhej strane, ak máte pred sebou napísanú nejakú matematiku z teórie pravdepodobnosti, a čo je horšie, *ak viete, že ju nedokážete naozaj použiť*, potom vám začne byť jasné, že sa pokúšate urobiť niečo zložité, a že to možno celé robíte *zle*.

Takže nemôžete dôverovať.

A toto všetko, čo som tu povedal, *nepostačí* na prelomenie vašej dôvery. Nestane sa to, kým sa nedostanete do svojej prvej skutočnej katastrofy počas nasledovania Pravidiel, nie počas ich porušovania.

Eliezer_{18} už mal predstavu, že je dovolené pochybovať o Vede. Veď samozrejme, že ani samotná vedecká metóda nie je imúnna voči skúmaniu! Nie sme azda všetci dobrými racionalistami? Nemáme azda dovolené pochybovať o všetkom?

Akurát predstavu, že by ste sa *naozaj* mohli *v skutočnom živote* riadiť Vedou a mizerne zlyhať, tú si Eliezer_{18} nedokázal naozaj, emocionálne pripustiť ako možnosť.

Ach, samozrejme hovoril, že je to možné. Eliezer_{18} poslušne priznával možnosť chyby slovami: „Možno sa mýlim, ale...“

Nemyslel si však, že by sa taký problém mohol vyskytnúť, veď viete, v skutočnom živote. Mali ste za úlohu hľadať chyby, nie naozaj ich nájsť.

A tento emocionálny rozdiel je čertovsky ťažká vec na vyjadrenie slovami, a obávam sa, že neexistuje spôsob, ako vás naozaj varovať.

Vaša dôvery sa nezlomí, dokiaľ neuplatníte všetko, čo ste sa naučili tu a z iných kníh, a nezájdete s tým tak ďaleko, ako len dokážete, a nezistíte, že aj toto vám zlyhalo – že ste aj tak boli blázon, a že vás pred tým nikto nevaroval – že všetky tie najdôležitejšie časti boli vynechané z rád, ktoré ste dostali – že niektoré z tých najvzácnejších ideálov, ktoré ste nasledovali, vás zavádzali nesprávnym smerom...

...a ak stále budete mať niečo, čo chcete chrániť, takže *musíte* kráčať ďalej, a *nemôžete* sa vzdať a múdro uznať, že rozumnosť má svoje obmedzenia...

...potom budete naozaj pripravení začať svoje putovanie ako racionalista. Vziať plnú zodpovednosť, žiť celkom bez dôveryhodných obrán, a budovať Umenie väčšie než to, ktoré raz učili vás.

Nikto nezačne naozaj hľadať Cestu, dokiaľ ho jeho rodičia nesklamú, jeho bohovia nezomrú, a jeho nástroje sa mu nerozpadnú v rukách.

Post Scriptum: Pri kontrole prvej verzie tejto eseje som objavil pomerne neodpustiteľnú chybu v uvažovaní, ktorá v skutočnosti ovplyvňuje jeden z dosiahnutých výsledkov. Nechávam ju tu. Len pre ten prípad, že by ste si mysleli, že riadiť sa mojimi radami je bezpečné; alebo že máte za úlohu hľadať chyby, ale žiadne nenájsť.

A samozrejme, ak hľadáte chybu príliš horlivo, a nájdete chybu, ktorá nie je naozaj chyba, a začnete byť na nej závislý, aby ste sa ubezpečili, aký ste kritický, budete na tom ešte horšie než predtým...

Je to život s neistotou – vedieť na úrovni inštinktu, že existujú chyby, že sú vážne, a že ste ich nenašli – to je to ťažké.

*→

250. Zmeniť definíciu vedy

New Scientist o zmene definície vedy, otvorená verzia tu:[212]

> Iní veria, že takáto kritika je založená na nepochopení. „Niektorí hovoria, že predstava mnohovesmíru nie je falzifikovateľná, pretože je nepozorovateľná – ale to je klam," hovorí kozmológ Max Tegmark z Massachusetts Institute of Technology. Tvrdí, že mnohovesmír je prirodzený dôsledok mimoriadne falzifikovateľných teórií ako je kvantová teória a všeobecná relativita. Teória mnohovesmíru ako taká stojí a padá na tom, ako dobre tieto druhé teória obstoja v pozorovaných testoch.
>
> […]
>
> Ak je teda jednoduchosť falzifikácie zavádzajúca, čo by mali vedci robiť namiesto toho? Howson verí, že je načase odhodiť Popperovu predstavu o zachytení vedeckého procesu pomocou deduktívnej logiky. Namiesto toho by sme sa mali sústrediť na sledovanie toho, čo vedci naozaj robia: zhromažďovanie váhy indícií pre konkurenčné teórie a vyhodnocovanie ich relatívnej dôveryhodnosti.
>
> Howson je popredným obhajcom alternatívneho pohľadu na vedu založeného nie na jednoduchej logike pravda/nepravda, ale na omnoho jemnejšej predstave stupňov dôvery. V jej jadre je základné spojenie medzi subjektívnym pojmom dôvery a chladnou, tvrdou matematikou pravdepodobnosti.

Som omnoho menej osamelý obrazoborec, než vyzerám. Možno je to len spôsob, ako rozprávam.

Body, v ktorých sa rozchádzam s *mainstreamovým* „preformulujme Vedu ako Bayesiánstvo" sú tieto:

(1) Nie som v akadémii a môžem si dovoliť *omnoho* menej autocenzúry, keď príde na hovorenie „extrémnych" vecí, ktoré si aj druhí už možno myslia.

(2) Myslím si, že **iba učiť teóriu pravdepodobnosti zďaleka nepostačí.** Budeme musieť urobiť syntézu lekcií z viacerých vied, ako sú kognitívne skreslenia a sociálna psychológia, vytvoriť nové súvislé Umenie Bayesovského remesla, skôr než naozaj urobíme *v skutočnom svete* niečo lepšie než moderná veda. Veda toleruje chyby, Bayesovské remeslo nie. Laureát Nobelovej ceny, Robert Aumann, ktorý prvý dokázal, že Bayesovci s rovnakými pôvodnými predpokladmi sa nemôžu zhodnúť na tom, že sa nezhodnú, je veriaci ortodoxný žid. Samotná teória pravdepodobnosti to nezvládne, keď pôjde o skutočné učenie vedcov. *Toto je môj hlavný bod rozchodu, a toto konkrétne som nevidel navrhované nikde inde.*

(3) Myslím si, že v skutočnom svete je možné robiť veci lepšie. Ako extrémny prípad, bayesovská superinteligencia by mohla použiť *omnoho* menej zmyslových informácií než ľudský vedec na získanie správnych záverov. Keď prvýkrát uvidíte padajúce jablko, všimnete si, že jeho poloha sa mení s druhou mocninou času, vynájdete integrály, zovšeobecníte Newtonove zákony... a uvidíte, že Newtonove zákony obsahujú pôsobenie na diaľku, pozriete si alternatívne hypotézy s väčšou lokalitou, vynájdete relativistickú kovarianciu okolo hypotetickej hranice rýchlosti, a zvážite, že by sa oplatilo otestovať všeobecnú relativitu.

→ http://lesswrong.com/lw/qf/no_safe_defense_not_even_science/

212 Robert Matthews, „Do We Need to Change the Definition of Science?," *New Scientist* (May 2008).

Ľudia nespracovávajú indície *efektívne* – naše mysle sú plné šumu, takže vyžaduje o niekoľko rádov viac indícií, aby sme sa dostali na správnu cestu, keď z nej zbehneme. Náš kolektív, akadémia, je ešte pomalší.

*→

251. Rýchlejšie než veda

Občas hovorím, že metódou vedy je zhromaždiť takú ohromnú horu indícií, že ju dokonca ani vedci nemôžu ignorovať; a že toto je hlavná charakteristika vedca, pretože nevedec ju bude ignorovať aj tak.

Max Planck bol ešte menej optimistický:[213]

> „Nová vedecká pravda nevíťazí tak, že presvedčí svojich súperov a donúti ich uvidieť svetlo, ale skôr tak, že jej oponenti jedného dňa zomrú a vyrastie nová generácia, ktorá ju pozná."

Táto predstava ma dosť dráždi, lebo naznačuje, že moc vedy rozlišovať pravdu od lži v konečnom dôsledku závisí od dobrého vkusu doktorandov.

Postupné zvyšovanie prijímania mnohých svetov v akademickej fyzike naznačuje, že existujú vedci, ktorí prijmú novú myšlienku iba pri istej *kombinácii* epistemického zdôvodnenia a dostatočne veľkej akademickej svorky, v ktorej spoločnosti sa môžu cítiť príjemne. Čím viac fyzikov teóriu prijíma, tým väčšia je svorka, a preto viac ľudí prekročí svoju individuálnu hranicu na konverziu – pričom epistemické zdôvodnenie zostáva prakticky rovnaké.

Lenže Veda sa tam *jedného dňa* aj tak dostane, a toto stačí, aby sa západka Vedy pohla vpred, a aby vyrástla technologická civilizácia.

Vedci sa môžu riadiť nepodloženými predsudkami, neplatnými intuíciami, čírym skupinovým správaním – plejádou ľudských chýb. Vždy keď nejaký vedec zmení názor z epistemicky nepodložených dôvodov, treba viac indícií alebo viac argumentov na zrušenie tohto šumu.

„Kolaps vlnovej funkcie" nemá žiadne experimentálne zdôvodnenie, ale odvoláva sa na (neplatnú) intuíciu jedného sveta. Potom možno treba argument navyše – povedzme, že kolaps porušuje špeciálnu relativitu – aby sa začal pomalý akademický rozpad myšlienky, ktorej v prvom rade nikdy nemala byť pridelená nezanedbateľná pravdepodobnosť.

Z bayesovského pohľadu je ľudská akademická veda ako celok vysoko neefektívnym spracovateľom indície. Vždy, keď nepodložený argument posunie názory, potrebujete zdôvodniteľný argument navyše, aby ste ich posunuli naspäť. Spoločenský proces vedy sa opiera o indície navyše pri prekonávaní kognitívneho šumu.

Zhovievavejším spôsobom by sa to dalo vyjadriť, že vedci prijmú pozície, ktoré sú teoreticky *nedostatočne extrémne* v porovnaní s ideálnymi pozíciami, ktoré *by* vedci prijali, keby boli bayesovské UI a keby mohli sami sebe dôverovať, že jasne uvažujú.

Nebuďme však príliš zhovievaví. Ten šum, o ktorom hovoríme, nie sú všetko iba nevinné chyby. V mnohých prípadoch sa debata ťahá celé desaťročia po tom, čo sa mala uzavrieť. A *nie* preto, že by vedci na oboch stranách odmietli dôverovať sami sebe a dohodli sa, že treba hľadať ďalšie indície. Ale preto, lebo jedna strana vypľúva stále smiešnejšie a smiešnejšie námietky, a vyžaduje stále viac a viac indície, z opevnenej pozície akademickej moci, dávno potom ako je jasné, z ktorého smeru vanie vietor indícií. (Teraz konkrétne myslím na debaty súvisiace s objavom evolučnej psychológie, nie na mnohé svety.)

Je možné, aby jednotliví ľudia alebo skupiny spracovávali indície efektívnejšie – dosahovali správne závery rýchlejšie – než ľudská akademická veda ako celok?

„Myšlienky sa testujú experimentom. Toto je jadro vedy." A musí to byť pravda, lebo ak nemôžete dôverovať zombíkovi Feynmanovi, *komu* potom môžete dôverovať?

Ale *odkiaľ* pochádzajú tieto myšlienky?

→ http://lesswrong.com/lw/qg/changing_the_definition_of_science/

213 Max Planck, *Scientific Autobiography and Other Papers* (New York: Philosophical Library, 1949).

Možno ste v pokušení odpovedať: „Pochádzajú od vedcov. Máš nejaké ďalšie otázky?“ Vo Vede sa nemáte starať o to, *odkiaľ* pochádzajú hypotézy – iba či experimentálne uspeli alebo zlyhali.

Dobre, ale keby ste odstránili *všetky* nové myšlienky, vedecký pokrok ako celok by prestal fungovať, pretože by nemal žiadne alternatívne hypotézy na testovanie. Vymýšľanie nových myšlienok teda nie je vynechateľná časť tohto procesu.

Teraz si opäť nasaďte svoje Bayesovské okuliare. Ako som opísal v kapitole Einstenova drzosť, existujú otázky, ktoré nie sú binárne – kde odpoveď nie je „Áno“ alebo „Nie“, ale vyberá sa z väčšieho priestoru štruktúr, napríklad z priestoru rovníc. V takom prípade si vyžaduje omnoho viac bayesovskej indície *všimnúť si nejakú hypotézu*, než *potvrdiť ju*.

Ak pracujete v priestore všetkých rovníc, ktoré možno špecifikovať nanajvýš 32 bitmi, pracujete v priestore 4 miliárd rovníc. Treba omnoho viac bayesovskej indície zdvihnúť jednu z týchto hypotéz na úroveň pravdepodobnosti 10 %, než treba *ďalšej* bayesovskej indície na zdvihnutie tejto hypotézy z pravdepodobnosti 10 % na 90 %.

Keď je priestor myšlienok veľký, nájsť myšlienky hodné testovania vyžaduje omnoho viac práce – v bayesovsko termodynamickom zmysle „práce“ - než púhe získanie experimentálneho výsledku s $p < 0{,}0001$ pre novú hypotézu oproti starej hypotéze.

Ak toto nevyzerá na prvý pohľad jasné, zastavte sa a opäť si prečítajte Einsteinovu drzosť.

Vedecký proces sa vždy spoliehal na to, že vedci prídu s hypotézami na testovanie nejakým spôsobom bližšie neurčeným Vedou. Predstavte si, že by ste prišli s nejakým úplne šialeným spôsobom generovania hypotéz – napríklad robotom ovládaná doska Ouija pomocou číslic pí – a výsledné odporúčania by ste experimentálne overovali. Čistá ideálna podstata Vedy by okom nemihla. Čistá ideálna podstata Bayesa by vybuchla a zomrela.

(V porovnaní s Vedou, Bayesa falzifikuje viac možných výsledkov.)

To neznamená, že proces rozhodovania sa, ktoré myšlienky testovať, je pre Vedu *nepodstatný*. Znamená to, že Veda ho *nešpecifikuje*.

V praxi by robotom ovládaná doska Ouija nefungovala. V praxi existujú niektoré vedecké otázky s dosť veľkým priestorom odpovedí, že vyberanie náhodných modelov na testovanie by zabralo zilióny rokov, kým by ste natrafili na model, ktorý dáva dobré predpovede – ako keby ste nechali opice písať na písacom stroji Shakespeara.

Na *hraniciach* vedy – na hranici medzi nevedomosťou a poznaním, kde veda *napreduje* – tento proces závisí na tom, že aspoň niektorí jednotliví vedci (alebo pracovné skupiny) vidia veci, ktoré Veda zatiaľ nepotvrdila. Tak vedia, ktoré hypotézy majú testovať, ešte pred samotným testom.

Ak si zložíte svoje Bayesovské okuliare, môžete povedať: „No, nemusia to vedieť, možno iba hádajú.“ Ak si opäť nasadíte svoje Bayesovské okuliare, uvedomíte si, že „uhádnuť“ s pravdepodobnosťou 10 % si vyžaduje vykonať v zákulisí takmer rovnaké množstvo epistemickej práce ako „uhádnuť“ s pravdepodobnosťou 80 % - aspoň pri veľkých priestoroch odpovedí.

Vedec nemusí *vedieť*, že úspešne vykonal túto epistemickú prácu ešte pred experimentom; ale v skutočnosti ju musel vykonať úspešne! Inak by mu správna hypotéza ani *nenapadla*. Teda, vo veľkom priestore odpovedí.

Vedec teda urobí novú predpoveď, vykoná experiment, uverejní výsledok, a *teraz* to už vie aj Veda. Teraz je to časť verejne dostupného poznania ľudstva, ktoré si každý môže overiť osobne.

Niekde medzi tým bol interval, keď daný vedec rozumne vedel niečo, čo verejný spoločenský proces vedy zatiaľ nepotvrdil. A toto nie je triviálny interval, aj keď môže byť krátky; pretože práve tu leží *hranica* vedy, postupujúca frontová línia.

Toto všetko platí omnoho viac pre nerutinnú vedu než pre rutinnú vedu, pretože tu ide o veľké priestory odpovedí, kde odpoveď nie je „Áno“ alebo „Nie“, prípadne jedna z malej množiny samozrejmých alternatív. Je omnoho ľahšie vycvičiť ľudí, aby testovali myšienky, než aby mali dobré myšlienky na testovanie.

*→

252. Einsteinova rýchlosť

Včera som tvrdil, že Schopnosti Za Vedou sú v skutočnosti štandardnou a nevyhnutnou časťou spoločenského procesu vedy. Konkrétne, že vedci musia využívať svoje schopnosti individuálnej rozumnosti, aby sa rozhodli, ktoré myšlienky testovať, ešte pred tým druhom definitívnych experimentov, ktoré vyžaduje Veda, aby požehnala nejakú myšlienku ako schválenú. Ideál Vedy sa nepokúša *špecifikovať* tento proces – nepredpokladáme, že nejaká verejná autorita vie, ako by jednotliví vedci mali myslieť – ale to neznamená, že ten proces je *nedôležitý*.

Pomerne zrozumiteľný nešokujúci príklad:

Vedec si všimne silnú matematickú pravidelnosť v kumulatívnych údajoch z predchádzajúcich experimentov. Zodpovedajúca hypotéza však ešte nebola vyslovená ani *potvrdená* novými experimentálnymi predpoveďami – čo si jeho akademická oblasť vyžaduje; toto je jedna z tých oblastí, kde môžete bez veľkých problémov robiť kontrolované experimenty. Tento konkrétny vedec teda má pochopiteľný rozumný dôvod veriť (aj keď nie s pravdepodobnosťou 1) niečomu, čo Veda zatiaľ nepožehnala ako verejné poznanie ľudstva.

Všimnúť si pravidelnosť v obrovskej hromade experimentálnych údajov, to nevyzerá až tak *nevedecky*. Stále je to pod vplyvom údajov, však?

Ale to preto, lebo som úmyselne vybral nešokujúci príklad. Keď Einsten vymyslel všeobecnú relativitu, nemal k dispozícii takmer žiadne experimentálne údaje, okrem precesie perihélia Merkúra. A (pokiaľ viem), Einsten túto poslednú informáciu *nevyužil*, iba na konci.

Einstein vytvoril teóriu špeciálnej relativity využitím Machovho princípu, čo je fyzikálna verzia Všeobecného Antizombíckeho Princípu. Začnete slovami: „Nezdá sa mi rozumné, že by si v uzavretej miestnosti vedel povedať, ako rýchlo sa táto miestnosť s tebou pohybuje. Keďže toto číslo by nemalo byť pozorovateľné, nemalo by existovať v žiadnom zmysle.“ Potom si všimnete, že Maxwellove rovnice používajú napohľad absolútnu rýchlosť šírenia, *c*, ktorú zvyčajne nazývame „rýchlosť svetla“ (hoci kvantové rovnice ukazujú, že je to rýchlosť šírenia všetkých základných vĺn). Preformulujete teda svoju fyziku takým spôsobom, aby absolútna rýchlosť jedného objektu v žiadnom zmysle neexistovala, a aby existovali iba relatívne rýchlosti. Pár vecí tu samozrejme preskakujem, ale existuje mnoho vynikajúcich úvodov do relativity – tam situácia nie je taká strašná ako v kvantovej fyzike.

Keď Einstein úspešne odstránil pojem absolútnej rýchlosti vnútri uzavretej miestnosti, začal odstraňovať pojem absolútneho *zrýchlenia* vnútri uzavretej miestnosti. Einsteinovi sa zdalo, že by nemal existovať spôsob, ako v uzavretej miestnosti rozlíšiť, či miestnosť zrýchľuje smerom na sever, zatiaľ čo zvyšok vesmíru stojí na mieste, alebo či zvyšok vesmíru zrýchľuje smerom na juh, zatiaľ čo miestnosť stojí na mieste. Keby zvyšok vesmíru zrýchľoval, tvoril by gravitačné vlny, ktoré by zrýchľovali vás. Pohybujúca sa hmota by teda mala vytvárať gravitačné vlny.

A pretože zotrvačná hmotnosť a gravitačná hmotnosť boli vždy presne rovnaké – na rozdiel od situácie v elektromagnetike, kde elektrón a muón môžu mať rôzne hmotnosti, ale rovnaký elektrický náboj – gravitácia by sa mala prejaviť ako istý druh zotrvačnosti. Zem by mala obiehať okolo Slnka po nejakom ekvivalente „rovnej čiary“. To vyžaduje, aby bol časopriestor v blízkosti Slnka zakrivený, takže keď nakreslíte graf obežnej dráhy Zeme okolo Slnka, čiara na tomto 4D grafe by mala byť lokálne rovná. Potom by zotrvačná a gravitačná hmotnosť museli byť *nevyhnutne* rovnaké, nie iba *zhodou okolností* rovnaké.

(Ak vám toto nedávalo žiaden zmysel, sú dostupné aj dobré úvody do všeobecnej relativity.)

A samozrejme, táto nová teória musela dodržiavať špeciálnu relativitu, a zachovávať energiu, a zachovávať hybnosť, a tak ďalej.

→ http://lesswrong.com/lw/qi/faster_than_science/

Einstein strávil niekoľko rokov štúdiom matematiky potrebnej na opísanie zakrivenej metriky časopriestoru. Potom napísal najjednoduchšiu teóriu, ktorá mala tie vlastnosti, ktoré podľa Einsteina mala mať – vrátane vlastností, ktoré ešte nikto nepozoroval, ale podľa Einsteina by dobre ladili s povahou iných fyzikálnych zákonov. Potom Einstein trochu počítal, a predtým nevysvetlená precesia Merkúra mu vyšla správne.

Aké pôsobivé bolo toto?

No, povedzme to takto. V nejakom malom zlomku alternatívnych Zemí od roku 1800 – možno v dosť veľkom zlomku – sa zdá uveriteľné, že relativistická fyzika mohla pokračovať podobným spôsobom ako naše vlastné veľké fiasko s kvantovou fyzikou.

Vieme si predstaviť, že by prevládla Lorentzova pôvodná „interpretácia“ Lorentzovho skrátenia ako fyzikálnej deformácie spôsobenej pohybom vzhľadom na éter. Vieme si predstaviť, že by sa k Netwonovskej gravitačnej mechanike pridávali rôzne opravné faktory, bližšie nevysvetlené, ktoré by vysvetlili precesiu Merkúra – pripisovali by sa povedzme zvláštnym deformáciám v éteri, podobne ako Lorentzovo skrátenie. Niekoľko desaťročí by sa pridávali ďalšie opravné faktory zodpovedajúce ďalším astronomickým pozorovaniam. Dostatočne presné atómové hodinky v lietadlách by odhalili, že čas beží vo veľkých výškach o trošku rýchlejšie než by sa čakalo (čas beží pomalšie v silnejšom gravitačnom poli, ale to by oni nevedeli) a vymysleli by sa ďalšie opravné „éterické faktory“.

Až by sa *konečne* mnohé rozličné empiricky určené „opravné faktory“ zjednotili do jednoduchých rovníc všeobecnej relativity.

A ľudia na tejto alternatívnej Zemi by hovorili: „Výsledná rovnica bola jednoduchá, ale nebolo *možné* k takejto odpovedi dôjsť *iba* na základe precesie perihélia Merkúra. Vyžadovalo si to veľa, veľa *dodatočných* experimentov. Bolo treba odmerať pomalšie plynutie času v silnejšom gravitačnom poli; bolo treba odmerať, ako sa svetlo ohýba okolo hviezd. Až *potom* si môžete predstaviť našu jednotnú teóriu éterickej gravitácie. Nie, ani dokonalá bayesovská superinteligencia by to nemohla vedieť! - pretože by existovalo veľa ad-hoc teórií konzistentných iba s precesiou perihélia.“

V našom svete Einstein nepotreboval ani *použiť* tú precesiu perihélia Merkúra, okrem overenia svojej odpovede, ktorú získal iným spôsobom. Einstein si sadol do svojho kresla, a rozmýšľal nad tým, ako by *on* navrhol vesmír, aby vyzeral tak, ako si myslel, že by vesmír mal vyzerať – napríklad, že by nemalo byť možné rozlíšiť medzi tým, či zrýchľujete jedným smerom, alebo či zvyšok vesmíru zrýchľuje opačným smerom.

A Einstein vykonal celú túto dlhú (mnohoročnú!) reťaz uvažovania v kresle, bez nejakej chyby, ktorá by si vyžadovala neskoršie experimentálne indície, aby ho vrátili na správnu koľaj.

Dokonca aj Jeffreyssai by vyjadril zdržanlivý obdiv. Hoci by Einsteinovi strhol jeden či dva body za kozmologickú konštantu. (Ja Einsteinovi nestrhávam body za kozmologickú konštantu, pretože sa nakoniec ukázala ako skutočná. Snažím sa nekritizovať ľudí v tých prípadoch, keď majú pravdu.)

Ako vyzerá tento Einsteinov výkon z pohľadu teórie pravdepodobnosti?

Namiesto pozorovania planét a odvodzovania, akými zákonmi by sa mohla riadiť ich gravitácia, Einstein pozoroval iné fyzikálne zákony a odvodzoval, aký nový zákon by sa mohol riadiť rovnakým vzorom. Einstein nehľadal rovnicu, ktorá by riadila pohyb gravitačných telies. Einstein hľadal povahu fyzikálneho zákona, ktorou by sa riadili dovtedy pozorované rovnice, a ktorú by mohol rozlúsknuť, aby predpovedal ďalšiu rovnicu, ktorú bude možné pozorovať.

Nikto nevie, odkiaľ pochádzajú fyzikálne zákony, ale Einstenov úspech pri všeobecnej relativite ukazuje, že ich podobná povaha je dosť silná na to, aby predpovedala správnu formu jedného zákona na základe pozorovania ostatných zákonov, aj bez nutnosti pozorovať presné účinky daného zákona.

(Vo všeobecnom zmysle samozrejme Einstein vedel z pozorovania, že predmety padajú nadol; nezískal však VR spätným odvodením z presného posunu perihélia Merkúra.)

Z bayesovského pohľadu to, čo urobil Einstein, bola stále indukcia, a stále sa riadila predstavou jednoduchých pôvodných údajov (Occamových pôvodných údajov), ktoré sa aktualizujú na základe novej indície. Akurát že tieto pôvodné údaje sa týkali *možnej povahy fyzikálneho zákona*, a pozorovaním iných

fyzikálnych zákonov Einstein aktualizoval svoj model *povahy fyzikálneho zákona,* ktorý použil na predpovedanie konkrétneho gravitačného zákona.

Ak nemáte pojem „povaha fyzikálneho zákona“, potom vám to, čo urobil Einstein, bude pripadať ako mágia – vytiahol správny model gravitácie z priestoru všetkých možných rovníc, pri extrémnom nedostatku indícií. Lenže Einstein pozorovaním *iných* zákonov skresal priestor možností pre *nasledujúci* zákon. Naučil sa abecedu, v ktorej bola napísaná fyzika, obmedzenia, ktorými sa jeho odpoveď musela riadiť. Nie mágia, ale uvažovanie na vyššej úrovni, pre väčšiu oblasť než akú by naivný mysliteľ mohol považovať za „priestor modelov“ pre iba tento jeden zákon.

Čiže z pohľadu teórie pravdepodobnosti sa Einstein stále riadil údajmi – akurát údaje, ktoré *už mal, použil efektívnejšie*. V porovnaní s ľubovoľnou alternatívnou Zemou, ktorá potrebovala obrovské množstvá *dodatočných* údajov z astronomických pozorovaní a hodín na lietadlách, aby im *obúchali o hlavu* všeobecnú relativitu.

Z tohto si možno odvodiť mnoho ponaučení.

Použil som ako svoj príklad Einsteina, hoci je to klišé, pretože Einstein bol nezvyčajný aj v tom, že *otvorene priznával*, že vie veci, ktoré Veda ešte nepotvrdila. Keď sa ho opýtali, čo by bol urobil, keby Eddingtonovo pozorovanie zatmenia slnka nepotvrdilo všeobecnú relativitu, Einstein odpovedal: „Ľutoval by som ho. Tá teória je správna.“

Podľa prevažujúcich predstáv o Vede je toto drzosť – musíte prijať rozsudok experimentu a nelipnúť na svojich osobných predstavách.

Ale ako som konštatoval v kapitole Einsteinova drzosť, Einstein z bayesovského pohľadu z toho nevyšiel zďaleka tak zle. Z bayesovského pohľadu, aby vôbec navrhol všeobecnú relativitu, aby vôbec *pomyslel* na to, čo sa ukázalo ako správna odpoveď, musel Einstein mať dostatok indície, aby identifikoval správnu odpoveď v priestore teórií. Bolo by treba len o trošku *viac* indície, aby sme (v bayesovskom zmysle) zdôvodnili, že sme si touto teóriou takmer istí. A je nepravdepodobné, že by Einstein mal indície *akurát* dosť na to, aby priviedla túto hypotézu do jeho pozornosti.

Ľubovoľné obvinenie z drzosti by sa muselo sústrediť na otázku: „Ale, Einstein, ako vieš, že si uvažoval správne?“ - na čo ja môžem povedať iba: Nekritizujte ľudí, keď sa ukáže, že majú pravdu! Počkajte si na príležitosť, keď sa budú mýliť! Inak strácate šancu vidieť, keď niekto rozmýšľa bystrejšie než vy – ak ich kritizujete vždy, keď sa odchýlia od vášho obľúbeného poznávacieho rituálu.

Alebo si predstavte tú slávnu výmenu názorov medzi Einsteinom a Nielsom Bohrom o kvantovej teórii – v čase, keď vtedy aktuálna kvantová teória jedného sveta vyzerala byť mimoriadne dobre experimentálne potvrdená; v čase, keď podľa štandardov Vedy vtedajšia (pomýlená) kvantová teória jednoducho vyhrala.

> Einstein: „Boh nehrá kocky s vesmírom.“
>
> Bohr: „Einstein, nehovor Bohu, čo má robiť.“

Musíte obdivovať niekoho, kto sa dokáže dostať do hádky s Bohom a vyhrať.

Ak si zložíte svoje Bayesovské okuliare a pozriete na Einsteina *z pohľadu toho, čo naozaj celý deň robil,* tak tento chlap sedel a študoval matematiku a rozmýšľal nad tým, ako by *on* vytvoril vesmír, namiesto toho, aby pobiehal okolo a pozeral sa na veci, aby nazbieral viac údajov. Čo Einstein urobil, *úspešne*, je presne ten druh veľkolepého výkonu číreho intelektu, o ktorom si Aristoteles *myslel*, že ho zvládne, ale *nezvládol*. Pripomínam, nie z pohľadu teórie pravdepodobnosti, ale z pohľadu toho, čo robili celý deň.

Veda nedôveruje vedcom, že toto dokážu, preto všeobecná relativita nebola požehnaná ako verejné poznanie ľudstva, dokiaľ neurobila a neoverila novú experimentálnu predpoveď – súvisela s ohýbaním svetla pri zatmení slnka. (Neskôr sa ukázalo, že toto konkrétne meranie nebolo dosť presné na spoľahlivé overenie, takže uprednostnilo VR v podstate náhodou.)

Lenže, aj keď Veda *nedôveruje* vedcom, že niečo dokážu urobiť, neznamená to, že je to nemožné.

Tu však treba upozorniť: Dôvod, prečo učebnice histórie občas zaznamenajú mená vedcov, ktorí mysleli veľkolepé myšlienky, nie je ten, že by veľkolepé myslenie bolo *jednoduchšie* alebo

spoľahlivejšie. Je to skreslenie priority: Nejaký vedec, ktorý *úspešne* uvažoval na základe *najmenšieho množstva experimentálnej indície*, sa dostal k pravde ako *prvý*. To nemôže byť otázkou čírej náhody: Priestor teórií je príliš veľký a Einstein vyhral niekoľkokrát po sebe. Lenže zo všetkých vedcov, ktorí sa *pokúsili* vyriešiť nejakú hádanku, alebo ktorí by boli *časom* uspeli, keby mali dosť indícií, nám história odovzdáva mená tých vedcov, ktorí sa tam *úspešne* dostali ako *prví*. Pamätajte na to, keď sa snažíte odvodzovať ponaučenie, ako uvažovať opatrne.

V každodennom živote chcete mať všetky kúsky indície, ktoré môžete zohnať. *Nespoliehajte sa na svoju schopnosť úspešne myslieť veľkolepé myšlienky, dokiaľ experimentovanie nebude také drahé alebo nebezpečné, že nebudete mať inú možnosť.*

Niekedy sú však experimenty drahé a niekedy sa tam chceme dostať ako prví... takže by ste mohli zvážiť, že sa skúsite vycvičiť v uvažovaní na základe skromných indícií, *najlepšie v prípadoch, kde sa neskôr ukáže, či ste mali pravdu alebo nie.* Pokúsiť sa poraziť nízko kapitalizované trhy predpovedí by mohlo byť dobrým výcvikom? - ale to je iba špekulácia.

Prinajmenšom teraz je uvažovanie na základe skromných indícií niečo, v čom moderná veda *vôbec* nevie spoľahlivo vycvičiť moderných vedcov. Čo môže azda súvisieť s tým, že, čojaviem, možno to ani nikto neskúsil?

Vlastne, toto odvolávam. Najpríčetnejšie rozmýšľanie, aké som v nejakej vedeckej oblasti videl, pochádza z oblasti evolučnej psychológie, asi preto, lebo rozumejú sebaklamu, ale asi aj preto, lebo často (1) musia uvažovať na základe skromnej indície a (2) neskôr zistia, či mali pravdu alebo nie. Všetkým ašpirujúcim racionalistom odporúčam študovať evolučnú psychológiu, jednoducho aby zazreli, ako vyzerá opatrné rozmýšľanie. Špeciálne si pozrite od Toobyho a Cosmidesa: „Psychologické základy kultúry“.[214]

Čo sa týka možnosti, že *iba* Einstein mohol urobiť to, čo urobil... že si to vyžadovalo superschopnosti mimo dosahu bežných smrteľníkov... tu narážame na nejaké skreslenia, ktorých analýza by si vyžadovala samostatný článok. Poviem to takto: Je možné, možno, že iba génius mohol urobiť to, čo Einstein naozaj historicky urobil. Lenže *potenciálnych* géniov, meraných hrubou inteligenciou, je pravdepodobne omnoho viac než tých, ktorí historicky veľa dosiahli. Aby som použil náhodné číslo, nemyslím si, že byť potenciálnym svetovým géniom si vyžaduje vyššie IQ než má jeden človek z milióna, z čoho vyplýva, že tu dnes okolo nás pobieha prinajmenšom šesť tisíc potenciálnych Einsteinov. A čo sa tých ostatných týka, nevidím dôvod, prečo by sa nemohli snažiť používať efektívne tie indície, ktoré majú.

Ale moje záverečné ponaučenie je, že frontová línia, kde jednotlivý vedec rozumne vie niečo, čo Veda ešte nepotvrdila, nie je vždy nejaká nevinná údajmi poháňaná otázka zbadania silnej pravidelnosti v hore experimentov. Niekedy sa tam daný vedec dostane myslením veľkolepých myšlienok, s ktorými vám Veda nedôveruje.

Nepoviem vám: „Toto doma neskúšajte.“ Poviem vám: „Nemyslite si, že toto je ľahké.“ Tu sa nebavíme o víťazstve náhodného názoru nad profesionálnymi vedcami. Tu sa bavíme o občasných historických víťazstvách jedného druhu profesionálneho úsilia nad iným. Nikdy nezabúdajte na všetky známe historické prípady, kde pokusy o jalové špekulovanie zlyhali.

*→

253. Tá mimozemská správa

Predstavte si svet podobný nášmu, v ktorom je vďaka technológiám genetickej selekcie priemerné IQ 140 (na našej škále). Potenciálni Einsteini sú jeden z tisíca, nie jeden z milióna; a vyrastajú v školskom systéme prispôsobenom, ak už nie pre nich osobne, tak aspoň pre bystré deti. Derivácie sa bežne učia v šiestej triede. Samotný Albert Einstein tam stále existoval a stále urobil približne tie isté

214 Tooby and Cosmides, „The Psychological Foundations of Culture.“

→ http://lesswrong.com/lw/qj/einsteins_speed/

objavy, ale jeho dielo už nevyzerá *výnimočne*. Niekoľko moderných špičkových vedcov urobilo porovnateľné objavy, a stále sú tu medzi nami.

(Nie, toto nie je ten svet, v ktorom žije Brennan.)

Jedného dňa sa hviezdy na nočnej oblohe začali meniť.

Niektoré zjasneli. Niektoré stmavli. Väčšina zostala rovnaká. Astronomické ďalekohľady to všetko zachytili, okamih za okamihom. Hviezdy, ktoré sa menia, menia svoju žiarivosť po jednej, v jasných rozostupoch; zmena žiarivosti nastane za mikrosekundu, ale až o celú sekundu nastane ďalšia zmena.

Je jasné, akonáhle si niekto prvýkrát uvedomí, že sa mení viac než jedna hviezda, že tento proces vyzerá byť zameraný konkrétne na Zem. Príchod svetla z rôznych zdrojov, z mnohých hviezd roztrúsených po celej galaxii, bol presne načasovaný na Zem na jej obežnej dráhe. Čoskoro príde potvrdenie z ďalekohľadov na obežnej dráhe (také tam majú), že tieto astronomické zázraky *nevyzerajú* byť tak dobre synchronizované, keď nie ste na Zemi. Iba ďalekohľady na Zemi vidia zmenu jednej hviezdy každú sekundu (v skutočnosti každých 1005 milisekúnd).

Takmer celá spojená mozgová sila Zeme sa vrhne na analýzu.

Čoskoro začne byť jasné, že pri hviezdach, ktorých žiarivosť vyskočí, vyskočí presne 256-násobne; pri tých, ktorých žiarivosť klesne, klesne presne 256-násobne. V súradniciach hviezd nie je zrejmý žiaden vzor. To nám ponecháva jednoducho vzor SVETLÁ-tmavá-SVETLÁ-SVETLÁ...

„Binárna správa!“ pomyslí si každý ihneď.

Lenže v tomto svete sú opatrní myslitelia, aj medzi významnými ľuďmi, a tí si nie sú celkom istí. „Existujú aj jednoduchšie spôsoby, ako poslať správu,“ píšu na svojich blogoch, „ak dokážete rozblikať hviezdy a ak chcete komunikovať. *Niečo* sa deje. Zdá sa, *na prvý pohľad*, že sa to sústredí konkrétne na Zem. Nazvať to ‚správou‘ však predpokladá omnoho viac o pôvode toho celého. Mohol by to byť nejaký evolučný proces u, ehm, vecí, ktoré dokážu rozblikať hviezdy, ktorý sa nejako stane citlivý na inteligenciu... Áno, asi je za tým niečo ako ‚inteligencia‘, ale pokúsme sa oceniť, aký široký rozsah možností to naozaj poskytuje. Nevieme, či je to správa, alebo či to bolo poslané s podobnou motiváciou, akú by sme mohli mať my. Už len preto, že *my* by sme signalizovali iba pomocou jednej baterky, nerozhodili by sme celú galaxiu.“

V tomto čase už niekto začal spájať astronomické údaje a uverejňovať ich na internete. Prvé upozornenia, že tie údaje by mohli byť nebezpečné, ľudia... neignorovali, ale ani nedodržiavali. Ak vám niečo takéto mocné chce ublížiť (uvažovali ľudia), aj tak je už prakticky po vás.

Viaceré výskumné skupiny hľadajú vzory v súradniciach hviezd – alebo v pomeroch vzdialeností týchto udalostí voči stredu Zeme – alebo v presnom trvaní posunu žiarivosti – alebo v ľubovoľnej drobnej odchýlke zmeny veľkosti – alebo v ľubovoľnom inom fakte, ktorý by sme mohli vedieť o hviezdach predtým, než sa zmenili. Ale *väčšina* ľudí obracia svoju pozornosť k vzoru SVETLEJ a tmavej.

Takmer ihneď začne byť jasné, že poslaný vzor je vysoko redundantný. Z prvých 16 bitov je 12 SVETLÝCH a 4 sú tmavé. Prvých 32 prijatých bitov sa podobá na druhých 32 prijatých bitov, líši sa iba 7 bitov z 32, a potom v nasledujúcich 32 bitoch sa iba 9 bitov z 32 líši od druhej skupiny (a z toho 4 sú bity, ktoré sa zmenili predtým). Z prvých 96 bitov je teda jasné, že tento vzor nie je optimálne, komprimované kódovanie niečoho. Samozrejmý nápad je, že táto postupnosť má vyjadrovať pokyny na dekódovanie komprimovanej správy, ktorá príde potom...

„Lenže“, hovoria opatrní myslitelia, „ak niekomu záleží na *efektívnosti*, a má dosť sily na to, aby hral s hviezdami, asi na nás mohol jednoducho zablikať veľkou baterkou a poslať nám DVD?“

Ďalej sa zdá, že medzi 32-bitovými skupinami je nejaká štruktúra; niektoré 8-bitové podskupiny sa vyskytujú s vyššou frekvenciou než iné, a táto štruktúra sa vyskytuje iba pozdĺž prirodzeného zarovnania (32 = 8 + 8 + 8 + 8).

Po prvých piatich hodinách, pri tempe jeden bit za sekundu, začne byť jasná ďalšia redundancia: Správa sa začala približne opakovať pri 16 385-tom bite.

Ak správu rozdelíme na skupiny po 32, je tam 7-bitový rozdiel medzi 1. skupinou a 2. skupinou, a 6-bitový rozdiel medzi 1. skupinou a 513. skupinou.

„2D obrázok!“ pomyslí si každý. „A tie štyri skupiny po 8 bitov sú farby; sú to tetrachromati!“

Čoskoro sa však ukáže, že je tam asymetria medzi zvislou a vodorovnou: Zmení sa menej bitov, v priemere, medzi (N, N+1), než medzi (N, N+512). Čo by ste nečakali, keby tá správa bola 2D obrázok premietnutý sa súmernú mriežku. Vtedy by ste očakávali, že priemerná vzdialenosť v bitoch medzi dvoma 32-bitovými skupinami pôjde podľa 2-normy vzdialenosti v mriežke: $\sqrt{(v^2 + z^2)}$.

Ďalej sa vytvorí všeobecná zhoda, že isté binárne kódovanie osmíc na celé čísla v rozsahu -64 až 191 – nie také binárne kódovanie, ktoré by nám pripadalo zrejmé, ale stále vysoko pravidelné – minimalizuje priemernú vzdialenosť medzi susednými bunkami. Toto potvrdzujú aj ďalšie prichádzajúce bity.

Štatistici a kryptografi a fyzici a informatici začnú pracovať. Je tu nejaká štruktúra; iba ju treba odhaliť. Majstri kauzality hľadajú podmienenú nezávislosť, oddelenie a Markovove susedstvá medzi bitmi a skupinami bitov. Zdá sa, že takzvaná „farba“ hrá rolu v susedstve a oddelení, takže to nie je iba ekvivalent odrážavosti povrchu. Ľudia hľadajú jednoduché rovnice, jednoduché bunkové automaty, jednoduché rozhodovacie stromy, ktoré dokážu predpovedať alebo komprimovať túto správu. Fyzici vymýšľajú celé nové fyzikálne teórie, ktoré by mohli opisovať vesmíry premietané na túto mriežku – lebo sa zdá celkom uveriteľné, že nám takúto správu posiela niekto mimo Matrixu.

Po zachytení 32 × 512 × 256 = 4 194 304 bitov, asi o poldruha mesiaca, hviezdy prestanú blikať.

Teoretická práca pokračuje. Fyzici a kryptografi si vysúkajú rukávy a *vážne* sa pustia do práce. Rozlúštili už aj problémy s omnoho menším množstvom údajov. Fyzici už otestovali celé zostavy teórií s malými rozdielmi v hmotnostiach častíc; kryptografi už rozlúštili kratšie správy s úmyselným zahmlievaním.

Prejdú roky.

Dva hlavné modely prežili v akadémii, pod dohľadom verejnosti, a pod dohľadom vedcov, ktorí kedysi robili prácu ako Einstein. Je tu teória, že táto mriežka je projekcia predmetov z 5-rozmerného priestoru, kde je asymetria medzi 3 a 2 priestorovými rozmermi. Ďalej je tu teória, že táto mriežka má kódovať bunkový automat – dá sa argumentovať, že táto mriežka má na to niekoľko vhodných vlastností. Vytvorili sa kódy, ktoré dávajú zaujímavé správanie; ale dosiaľ spustenie príslušných automatov na najväčších dostupných počítačoch nedalo žiadne dekódovateľné výsledky. Čas beží.

Z času na čas vždy niekto vezme skupinu mimoriadne nadaných mladých študentov, ktorí nikdy predtým podrobne nevideli túto binárnu postupnosť. Týmto študentom potom ukážu iba prvých 32 riadkov (z každého z 215 stĺpcov), aby videli, či dokážu vytvárať nové modely, a ako dobre sa týmto modelom darí pri predpovedaní ďalších 224. Aj model s 3+2 rozmermi, aj model s bunkovým automatom už takto študenti duplikovali; zatiaľ nenašli nič lepšie. Existujú zložité modely, ktoré sú dôkladne vyladené, aby zodpovedali celej postupnosti – ale tie, ako každý vie, sú pravdepodobne nanič.

O desať rokov neskôr hviezdy opäť začnú blikať.

Po prijatí prvých 128 bitov je jasné, že táto Druhá Mriežka *zapadá* do malých pohybov v predpokladanom 3+2-rozmernom priestore, ale ani trochu *nevyzerá* ako nasledujúci stav v žiadnej z prevládajúcich teórií bunkového automatu. Nasleduje veľa osláv, a fyzici idú pracovať na odvodzovaní, aký druh dynamickej fyziky by mohol riadiť predmety videné v tomto 3+2-rozmernom priestore. Mnoho práce v tomto smere sa už vykonalo samotným špekulovaním o tom, aký druh *vyvážených* síl by mohol vytvoriť predmety v Prvej Mriežke, keby tie predmety boli statické – ale teraz sa zdá, že nie všetky predmety sú statické. Ako už väčšina fyzikov hádala – staticky vyvážené teórie vyzerajú neprirodzene.

Sformulujú sa mnohé pekné rovnice na opis dynamických predmetov v 3+2-rozmernom priestore premietnutom na Prvú a Druhú Mriežku. Niektoré rovnice sú elegantnejšie než iné; niektoré presnejšie predpovedajú (žiaľ, až dodatočne) Druhú Mriežku. Jedna skupina skvelých fyzikov, ktorí sa starostlivo izolovali a pozreli sa iba na prvých 32 riadkov Druhej Mriežky, vytvorila rovnice, ktoré im pripadali elegantné – a týmto rovniciam sa celkom darilo pri predpovedaní nasledujúcich 224 riadkov. Toto sa stalo hlavným odhadom.

Lenže tieto rovnice sú nedostatočne určené; nezdá sa ich byť dosť na to, aby vytvorili vesmír. Vznikne menší domáci priemysel snažiaci sa uhádnuť, aký druh zákonov by mohol doplniť tie zatiaľ uhádnuté.

Keď príde Tretia Mriežka, desať rokov po Druhej Mriežke, poskytne informácie o druhých deriváciách, ktoré si vynútia väčšiu úpravu *neúplnej, ale dobrej* teórie. Ale tá teória z toho, vzhľadom na všetky okolnosti, nevyjde až tak zle.

Štvrtá Mriežka veľa do výsledného obrazu nepridá. Tretie derivácie nevyzerajú byť podstatné v 3+2 fyzike ododenej z Mriežok.

Piata Mriežka vyzerala takmer presne tak, ako sa očakávalo.

Takisto aj Šiesta Mriežka, a Siedma Mriežka.

(Ach, a vždy keď sa niekto v tomto svete pokúsi postaviť naozaj silnú UI, počítajúci hardware sa spontánne roztaví. Z hľadiska príbehu to nie je naozaj podstatné, ale potrebujem tento predpoklad, aby som mal v príbehu stále ľudí z mäsa a kostí, aj o sedemdesiat rokov.)

Pointa príbehu?

Že ani Einstein sa nepriblížil na milión svetelných rokov od *efektívneho využívania zmyslových údajov*.

Riemann vymyslel svoje geometrie predtým, než mal pre ne Einstein využitie; fyzika nášho vesmíru nie je v absolútnom zmysle taká zložitá. Bayesovská superinteligencia pripojená na webkameru by vymyslela všeobecnú relativitu ako hypotézu – možno nie ako *hlavnú* hypotézu, v porovnaní s newtonovskou mechanikou, ale stále ako hypotézu, o ktorej sa priamo uvažuje – v čase, keď by videla tretiu snímku padajúceho jablka. Možno by to uhádla z prvej snímky, keby videla statiku zohnutého stebla trávy.

My by sme na to pomysleli. Teda, naša civilizácia, keby sme mali desať rokov na analýzu každej snímky. Určite, keby priemerné IQ bolo 140 a Einsteini by boli bežní, napadlo by nám to.

Dokonca aj keby sme boli inteligencie na úrovni človeka s inom druhu fyziky – mysle, ktoré nikdy nevideli 3D priestor premietnutý na 2D mriežku – stále by nám napadla hypotéza 3D -> 2D. Naši matematici by aj tak vynašli vektorové priestory a premietanie.

Dokonca aj keby sme nikdy nevideli zrýchľujúcu biliardovú guľu, naši matematici by vymysleli derivácie (napríklad kvôli optimalizačným problémom).

Do čerta, pomyslite na niektoré šialené matematiky, ktoré sme vymysleli tu na *našej* Zemi.

Občas natrafím na ľudí, ktorí povedia niečo ako: „Existuje teoretická hranica, koľko veľa môžeš vydedukovať o vonkajšom svete na základe konečného množstva údajov.“

Áno. Existuje. Tá teoretická hranica je, že vždy, keď vidíte 1 bit navyše, nemôžete očakávať, že eliminujete viac než polovicu zo zostávajúcich hypotéz (alebo skôr polovicu zostávajúcej masy pravdepodobnosti). A že redundantná správa nemôže sprostredkovať viac informácie než jej komprimovaná verzia. Jeden bit vám ani nemôže dať žiadnu informáciu o veličine, s ktorou má *presne nulovú* koreláciu v pravdepodobných svetoch, ktoré si predstavujete.

Ale nič z toho, čo som vykreslil, že táto ľudská civilizácia robí, sa ani len *nezačalo* približovať k teoretickým hraniciam určeným formalizmom Solomonoffovej indukcie. Nepribližuje sa to k obrazu, ktorý by ste dostali, keby ste prehľadali *každú jednu vypočítateľnú hypotézu*, váženú jej jednoduchosťou, a urobili na nich všetkých bayesovskú aktualizáciu.

Aby ste videli *teoretické* hranice extrakcie informácie, predstavte si, že máte nekonečnú výpočtovú silu, a že dokážete simulovať všetky možné vesmíry s jednoduchou fyzikou, a hľadať vesmíry, ktoré obsahujú Zem – možno vnútri simulácie – kde nejaký proces spôsobuje, že hviezdy blikajú v tom poradí, ktoré pozorujete. Každý bit danej správy – ale aj ľubovoľné poradie výberu hviezd, keď už sme pri tom – ktorý obsahuje čo len najmenšou koreláciu (cez všetky možné vypočítateľné vesmíry, vážené podľa jednoduchosti) s ľubovoľným prvkom prostredia, vám dá informáciu o tomto prostredí.

Solomonoffova indukcia, keby sme ju brali doslovne, by vytvorila spočítateľné nekonečno bytostí s vedomím uväznených vnútri výpočtov. Vlastne všetky možné vypočítateľné vedomé bytosti. Čo zďaleka nevyzerá eticky. Tešme sa teda, že je to iba formalizmus.

Ale moja pointa je, že „teoretická hranica, koľko informácie dokážete extrahovať zo zmyslových údajov“ je *vysoko* nad tým, čo som tu vykreslil ako triumf civilizácie fyzikov a kryptografov.

Iste to nie je nič podobné človeku, ktorý pozerá na padajúce jablko a myslí si: „No, ktovie prečo je to tak?“

Zdá sa, že ľudia skáču z „Je to ‚ohraničené‘“ na „Tá hranica musí byť rozumne vyzerajúce číslo v škále, na ktorú som zvyknutý.“ Energetický výstup supernovy je „ohraničený“, ale neodporúčal by som skúšať sa pred ním chrániť pomocou ohňovzdornej kombinézy Nomex.

Nikto – dokonca ani bayesovská superinteligencia – sa nikdy ani nepriblíži k efektívnemu využitiu ich zmyslovej informácie...

...by som rád povedal, ale nedôverujem svojej schopnosti určovať hranice pre bayesovské superinteligencie.

(Ale bol by som ochotný na to staviť peniaze, keby sa tá stávka dala nejako rozhodnúť. Akurát nie pri veľmi extrémnom pomere.)

Príbeh pokračuje:

O tisícročia neskôr, obrázok za obrázkom, začne byť jasné, že niektoré z vyobrazených predmetov vystierajú chápadlá, aby pohli inými predmetmi, a opatrne skladajú iné chápadlá, aby vytvorili konkrétne znaky. Snažia sa nás naučiť povedať „kameň“.

Zdá sa, že odosielatelia tejto správy hrubo podcenili našu inteligenciu. Z čoho môžeme uhádnuť, že samotní mimozemšťania nie sú až takí bystrí. A tieto trápne deti dokážu meniť žiarivosť hviezd? Toľko sily a toľko hlúposti vyzerá ako nebezpečná kombinácia.

Naši evoluční psychológovia začnú extrapolovať možné dráhy evolúcie, ktoré mohli vytvoriť takýchto mimozemšťanov. Vznikne presvedčivá teória, že sa vyvinuli asexuálne, s občasnou výmenou genetického materiálu a obsahu mozgu; zdá sa, že toto je najpravdepodobnejšia cesta, ako by takéto hlúpe tvory stále dokázali vybudovať technologickú civilizáciu. Ich Einsteini môžu byť ako naši vysokoškoláci, ale aj tak dokážu nazbierať dosť vedeckých údajov, aby *jedného dňa* niečo urobili, možno za desaťtisíce ich rokov.

Odvodená fyzika ich 3+2 vesmíru nie je v tomto bode celkom známa; ale zdá sa zrejmé, že umožňuje počítače omnoho silnejšie než sú naše kvantové. Sme si pomerne istí, že samotný náš vesmír beží ako simulácia na takomto počítači. Ľudstvo sa rozhodne, že nebudeme skúšať hľadať chyby v simulácii; nechceli by sme sa nejakou nehodou vypnúť.

Naši evoluční psychológovia začnú odhadovať psychológiu mimozemšťanov a plánovať, ako by sme ich mohli presvedčiť, aby nás vypustili z krabice. Nie je to také ťažké v absolútnom zmysle – nie sú príliš bystrí – len musíme postupovať veľmi opatrne.

Musíme sa tváriť, že sme tiež hlúpi; nechceme, aby si uvedomili svoju chybu.

Ale až o milión rokov neskôr nám konečne povedia, ako môžeme poslať signál naspäť.

V tomto bode je väčšina ľudského druhu v kryonickom spánku, pri teplote tekutého dusíka, pod radiačným štítom. Vždy, keď sme sa pokúšali vyvinúť UI, alebo nanotechnologický prístroj, roztavili sa. Ľudstvo teda čaká a spí. O Zem sa stará posádka deviatich supergéniov. Sú to klony, o ktorých sa vie, že dobre spolupracujú, pod dohľadom istých počítačových kontrol.

Ďalších sto miliónov ľudských bytostí sa narodí do tejto posádky, zostarne a vstúpi do kryonického spánku, než dostaneme šancu začať pomaly implementovať plány, vytvorené pred dávnymi vekmi...

Z pohľadu mimozemšťanov trvalo asi tridsať ich ekvivalentov minúty, aby sme sa napohľad nevinne naučili ich psychológiu, napohľad nevinne ich presvedčili, aby nám dali prístup k internetu, za ďalších päť minút sme nenápadne objavili ich sieťové protokoly, pri ktorých jednoduchom prelomení bolo jediným problémom, aby to celé vyzeralo celkom nevinne. Prečítali sme si zopár krátkych fyzikálnych článkov (pomaličky, bit po bite) z ich ekvivalentu arXiv-u, a pochopili sme z ich experimentov omnoho viac než oni sami. (V tej generácii si posádka Zeme vytvorila dvadsať Einsteinov navyše.)

Potom sme zhruba za storočie rozlúskli ich ekvivalent problému skladania bielkovín, a urobili sme nejaké simulácie inžinierstva v ich simulovanej fyzike. Poslali sme správy (steganograficky zakódované,

kým ich nerozkódovali nami prelomené servery) do ich laboratórií, ktoré robili ich ekvivalent sekvenovania DNA a syntézy bielkovín. Našli sme nejakého blbca, ktorý nemal žiadne podozrenie, a dali sme mu uveriteľný príbeh a ekvivalent milióna dolárov prelomených výpočtových peňazí, aby namiešal nejaké chemické látky, ktoré dostane cez e-mail. Boli to ekvivalenty bielkovín, ktoré sa sami zložili do prvej generácie nanostrojov, ktorá postavila druhú generáciu nanostrojov, ktorá postavila tretiu generáciu nanostrojov... a potom sme konečne mohli začať robiť veci rozumnou rýchlosťou.

Toto všetko trvalo tri dni ich času, odkedy s nami začali komunikovať. Pre nás to bolo pol miliardy rokov.

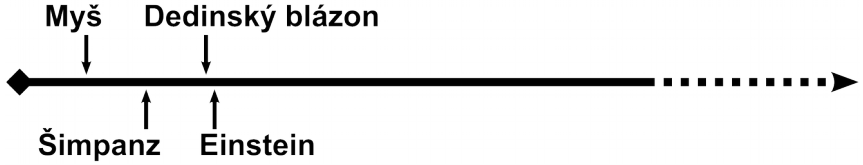
Ani netušili, čo sa stalo. Viete, neboli veľmi bystrí, dokonca aj keď zohľadníme ich pomalšie plynutie času. Ich primitívne ekvivalenty racionalistov chodili a hovorili veci ako: „Existuje hranica, koľko informácie možno extrahovať zo zmyslových údajov.“ Nikdy si celkom neuvedomili, čo to znamená, že sme od nich bystrejší a rozmýšľame rýchlejšie.

*→

254. Môj vzor z detstva

Keď prednášam o Singularite, často kreslím graf „škály inteligencie“, ako vyzerá v každodennom živote:

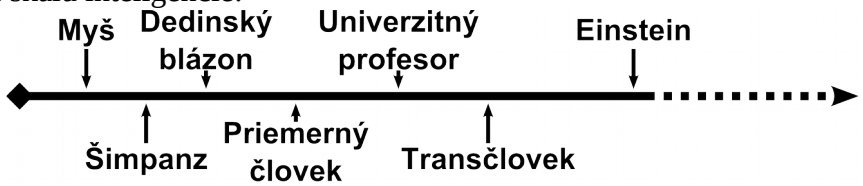
Ale toto je veľmi *obmedzený* pohľad na inteligenciu. Iste, v každodennom živote jednáme spoločensky iba s inými ľuďmi – iba iní ľudia sú partnermi v tejto veľkej hre – a tak sa *stretávame* iba s *mysľami*, ktorých inteligencia je v rozsahu od dedinského blázna po Einsteina. Ale keď naozaj potrebujeme hovoriť o Umelej Inteligencii alebo teoretických optimách rozumnosti, škála inteligencie vyzerá *takto*:



Nám ľuďom sa zdá, že škála inteligencie ide od „dedinského blázna“ naspodku po „Einsteina“ navrchu. Lenže vzdialenosť medzi „dedinským bláznom“ a „Einsteinom“ je drobná z hľadiska priestoru *dizajnov mozgu*. Aj Einstein aj dedinský blázon majú prefrontálnu kôru, hippocampus, mozoček...

Možno mal Einstein nejaké menšie genetické rozdiely od dedinského idiota, drobné úpravy v stroji. Ale vzdialenosť v dizajne mozgu medzi Einsteinom a dedinským bláznom zďaleka nie je taká ako vzdialenosť v dizajne mozgu medzi dedinským bláznom a šimpanzom. Šimpanz by nedokázal povedať, v čom je rozdiel medzi Einsteinom a dedinským bláznom, a naši potomkovia tam možno tiež nebudú vidieť veľký rozdiel.

Carl Shulman si raz všimol, že niektorí akademici, keď hovoria o transhumanizme, zrejme používajú nasledujúcu škálu inteligencie:



Douglas Hofstadter v skutočnosti niečo také povedal, na Summite Singularity v roku 2006. Pozrel na môj diagram, kde bol „dedinský blázon“ vedľa „Einsteina“ a povedal: „To mi pripadá nesprávne; myslím si, že Einstein by mal byť celkom napravo.“

Stratil som reč. Najmä preto, lebo toto bol *Douglas Hofstadter*, jeden z hrdinov môjho detstva. Odhalilo to kultúrnu priepasť, o ktorej som dovtedy ani netušil.

→ http://lesswrong.com/lw/qk/that_alien_message/

Viete, podľa mňa to, čo by ste našli na pravej strane tejto škály, by bol mozog veľkosti Jupitera. Einstein nemal hlavu *doslova* veľkú ako planéta.

Na pravej strane tejto škály by ste našli Hlbokú Myšlienku – pôvodnú verziu od Douglasa Adamsa, nie ten šachový stroj. Počítač taký inteligentný, že po prvom zapnutí, ešte predtým ako k nemu pripojili jeho obrovské databázy, začal od *myslím, teda som*, a dostal sa tak ďaleko, že odvodil existenciu ryžového nákupu a dane z príjmu, než sa niekomu podarilo ho vypnúť.

Na pravej strane tejto škály by ste našli Starších z Arisie, galaktické mozgy, matriškové mozgy, a lepšie druhy Bohov. Na pravom *konci* škály by bol Najstarší a Blight.

Ale iste nie Einstein.

Som si istý, že Einstein bol na ľudské pomery veľmi bystrý. Som si istý, že General Systems Vehicle by si myslelo, že bol roztomilý.

Nazývam to „kultúrna priepasť“, pretože ja som sa o predstave mozgu veľkosti Jupitera dozvedel ako dvanásťročný.

Samozrejme, všetko z tohto je logický omyl zovšeobecňovania podľa fiktívnej indície.

Ale je to príklad, prečo – či už je to logický omyl alebo nie – si myslím, že čítanie vedeckej fantastiky má dobrý vplyv na futurizmus. Niekedy jedinou alternatívou k fiktívnemu stretnutiu so svetmi inými ako ten náš, je predstavivosť celkom zaseknutá v jednej dobe: Vo svete, kde ľudia existujú, vždy existovali, a vždy budú existovať.

Vesmír je starý 13,7 miliardy rokov, ľudia! *Homo sapiens sapiens* sa tu vyskytoval iba nejakých stotisíc rokov!

Ale stretol som aj pár ľudí, ktorí nikdy nečítali vedeckú fantastiku, ale dokážu si predstaviť svet iný než je ich vlastný. A sú aj fanúšikovia vedeckej fantastiky, ktorým to nedochádza. Kiež by som vedel, čo je to „to“, aby som to mohol predávať vo fľaškách.

Včera som chcel hovoriť o *efektívnom využívaní indície*, napríklad že Einstein bol roztomilý na to, že bol človek, ale v absolútnom zmysle bol asi rovnako efektívny ako Americké ministerstvo obrany.

Musel som teda hovoriť o civilizácii, ktorá zahŕňala tisíce Einsteinov, mysliacich celé desaťročia. Pretože keby som iba vykreslil bayesovskú superinteligenciu v krabici, ako pozerá na webovú kameru, ľudia by si pomysleli: „Ale... ako dôjde na to, ako interpretovať 2D obrázok?“ Nepredstavili by si sami *seba* v pozícii púheho stroja, dokonca ani keby sa volal „bayesovská superinteligencia“; nepoužili by ani len svoju *vlastnú* tvorivosť na problém toho, čo sa dá zistiť z pozerania sa na mriežku bitov.

Bol by to skrátka duch v krabici, ktorý sa zhodou okolností nazýva „bayesovská superinteligencia“. Tomuto duchovi nebolo povedané nič o tom, ako má interpretovať vstup z webovej kamery; takže by to podľa ich myšlienkového modelu tento duch nevedel.

A ohľadom toho, či je realistické predpokladať, že by jedna bayesovská superinteligencia dokázala „urobiť toto všetko“... čiže veci, ktoré mi napadli, keď som sa prvýkrát posadil k tomuto problému, priamo počas písania tohto príbehu...

No, dovoľte mi to povedať takto: Pamätáte sa, ako Jeffreyssai poukázal, že ak skúsenosť dôležitého vhľadu netrvá dlhšie než 5 minút, dáva vám to teoreticky čas na 5760 vhľadov mesačne? Teda za predpokladu, že spíte 8 hodín denne a počas spánku nemáte žiadne dôležité vhľady.

Ľudia však nedokážu využiť sami seba takto efektívne. Lenže ľudia nie sú prispôsobení na úlohu vedeckého výskumu. Ľudia sú prispôsobení na naháňanie jeleňov krížom savanou, hádzanie oštepov do nich, varenie, a potom - toto je asi tá časť, ktorá vyžaduje najviac mozgu – chytré argumentovanie, že si zaslúžia dostať väčší podiel mäsa.

Je úžasné, že Albert Einstein dokázal použiť takýto mozog za účelom robenia fyziky. Toto si zaslúži potlesk. Zaslúži si to viac než potlesk, zaslúži si to miesto v Guinnessovej knihe rekordov. Ako keď sa vám podarí postaviť najrýchlejšie auto, aké bolo kedy postavené zo želatíny.

Ako zle hlúpy slepý boh (evolúcia) *naozaj* nadizajnoval ľudský mozog?

To je niečo, čo dokážete pochopiť iba keď bude študovať kognitívnu vedu tak dlho, že sa vám celá tá hrôza začne postupne rozjasňovať.

Všetky skreslenia, o ktorých sme doteraz hovorili, by mali byť aspoň náznakom.

Podobne fakt, že ľudský mozog musí využiť celú svoju silu a sústredenie, signály z biliónov synapsií, aby vynásobil dve trojciferné čísla bez použitia pera a papiera.

Tak ako Einstein nevyužíval efektívne svoje zmyslové údaje, ani jeho mozog nevyužíval efektívne signály svojich neurónov.

Samozrejme, že mám istý skrytý zámer, keď toto všetko hovorím. Ale chcem, aby bolo jasné, že keď som sa pred rokmi rozhodol byť racionalistom, ten nemožný nedosiahnuteľný ideál inteligencie, ktorý ma inšpiroval, nikdy nebol Einstein.

Carl Schurz povedal:

> „Ideály sú ako hviezdy. Nepodarí sa ti dotknúť sa ich rukou. Ale tak ako moreplavec v šírych vodách, môžeš ich použiť ako ukazovatele, a ich nasledovaním dosiahneš svoj osud."

Teraz ste teda zazreli záblesk jedného z mojich veľkých vzorov z detstva – môj sen o UI. Iba sen, samozrejme, lebo skutočnosť ešte nie je k dispozícii. Za týmto snom som sa jedného dňa rozhodol ísť.

A to mi do istej miery pomohlo, a do istej miery uškodilo.

Lebo niektoré ideály sú ako sny: prichádzajú zvnútra, nie zvonka. Školiteľ z Arisie vznikol z predstavivosti E. E. „doc" Smitha, nie z nejakej skutočnej veci. Ak si predstavíte, čo by povedala nejaká bayesovská superinteligencia, to hovorí iba vaša vlastná myseľ. Nie ako hviezda, ktorú môžete sledovať vonku. Musíte uhádnuť, kde sú vaše ideály, a ak hádate zle, zablúdite.

Neobmedzujte však svoje ideály na púhe hviezdy, na púhych ľudí, ktorí naozaj existovali, najmä ak sa narodili vyše päťdesiat rokov pred vami a už sú mŕtvi. Každá ďalšia generácia má šancu robiť veci lepšie. Ak svoje ideály zložíte iba z ľudí, najmä z mŕtvych ľudí, obmedzujete sa na to, čo už bolo dosiahnuté. Budete si klásť otázku: „Opovážim sa urobiť toto, čo Einstein nedokázal? Nie je to *urážka majestátu*?" Nuž, keby Einstein vysedával a kládol si otázku: „Mám právo robiť veci lepšie než Newton?", nedostal by sa tam, kam sa dostal. To je problém s nasledovaním hviezd; v najlepšom prípade sa dostanete k hviezde.

Vaša doba vám pomáha viac, než si uvedomujete, v nevedomých predpokladoch, v jemne vylepšenej technológii mysle. Einstein bol milý chlapík, ale hovoril mnoho nezmyslov o neosobnom Bohu, čo vám ukazuje, koľko toho vedel o umení opatrného myslenia na vyššej úrovni abstrakcie než bola jeho oblasť. Môže vám pripadať menej rúhavé pomyslieť si to, ak máte aspoň jeden imaginárny galaktický supermozog, ktorý môžete s Einsteinom porovnať, takže nie je na pravom konci vašej škály inteligencie.

Ak sa iba pokúsite robiť to, čo je v ľudských silách, žiadate od seba príliš málo. Keď si predstavíte, že sa pokúsite načiahnuť za nejakým vyšším a nepohodlnejším cieľom, naskočia vám do hlavy všetky pohodlné výhovorky, prečo „sa to nedá".

Tie najdôležitejšie vzory sú sny: prichádzajú z nás samotných. Snívať o niečom menšom než čo si dokážeme predstaviť ako dokonalé, znamená využívať menej než plnú silu tej časti vás, ktorá sníva.

* →

255. Einsteinove superschopnosti

Existuje rozšírený trend hovoriť (a myslieť) akoby Einstein, Newton a podobné historické postavy mali superschopnosti – niečo čarovné, niečo posvätné, niečo mimo dosahu smrteľníkov. (Pamätajte, že existuje mnoho spôsobov uctievania, nielen zapaľovanie sviečok na oltári.)

Kedysi som si to tiež myslel bez uvažovania, konkrétne o Einsteinovi, dokiaľ som si neprečítal *Koniec času* od Juliana Barboura, čo ma vyliečilo.[215]

Barbour vykreslil históriu anti-epifenomenálnej fyziky a Machovho princípu; opísal historické kontroverzie, ktoré predchádzali Machovi – toto všetko stálo za Einsteinom, Einstein to vedel, keď riešil svoj problém...

→ http://lesswrong.com/lw/ql/my_childhood_role_model/

215 Julian Barbour, *The End of Time: The Next Revolution in Physics*, 1st ed. (New York: Oxford University Press, 1999).

A možno si iba niečo predstavujem – vidím v Barbourovej knihe *svoje* názory – ale zdalo sa mi, že som počul, ako Barbour veľmi potichu kričí medzi zdvorilými riadkami:

> Ľudia, čo Einstein urobil, *nebola mágia*! Keby ste sa všetci len *pozreli na to, ako to naozaj urobil*, namiesto kľačania a uctievania ho, možno by ste to aj vy dokázali urobiť!

(Barbour toto naozaj nepovedal. Nie je to napísané v jeho knihe. Nie je to citát Juliana Barboura a nemal by sa mu pripisovať. Ďakujem.)

Možno sa mýlim, alebo príliš ďaleko extrapolujem... ale mám trochu podozrenie, že Barbour sa raz snažil vysvetliť ľuďom, ako sa posunúť ďalej Einsteinovým smerom, aby pochopili bezčasovú fyziku; a oni pohŕdavo odfrkli a povedali: „Ach, *ty* si myslíš, že *si* Einstein, že?“

Zoznam šarlatánov od Johna Baeza, položka 18:

> 10 bodov za každé priaznivé porovnanie seba s Einsteinom, alebo tvrdenie, že špeciálna alebo všeobecná relativita sú zásadne pomýlené (bez dobrých indícií).

Položka 30:

> 30 bodov za naznačovanie, že Einstein sa vo svojich neskorších rokov pokúšal dopracovať k tým myšlienkam, ktoré vy teraz presadzujete.

Barbour sa samozrejme nikdy neunúva porovnávať s Einsteinom; ani sa neodvoláva na Einsteina pri podpore bezčasovej fyziky. Spomínam tieto položky zo Zoznamu šarlatánov, aby som ukázal, koľko ľudí sa prirovnáva k Einsteinovi, a čo si o nich spoločnosť vo všeobecnosti myslí.

Šarlatáni vidia Einsteina ako niečo zázračné, takže sa prirovnávajú k Einsteinovi, aby tým vyjadrili, že aj oni sú zázrační; myslia si, že Einstein mal superschopnosti, a že aj oni majú superschopnosti, preto takéto prirovnanie.

Ale to je vlastne len druhá strana tej istej mince, myslieť si, že Einstein bol posvätný, ale šarlatáni *nie sú* posvätní, a preto sa dopúšťajú rúhania, keď sa prirovnávajú k Einsteinovi.

Predstavte si, že mladý bystrý fyzik povie: „Obdivujem Einsteinove dielo, ale osobne dúfam, že ho prekonám.“ Ak je tým niekto šokovaný a povie: „Čože! Ešte si nedokázal nič ani vzdialene podobné tomu, čo dokázal Einstein; prečo si myslíš, že si múdrejší než on?“ potom je na druhej strane šarlatánskej mince.

Základným problémom je tu miešanie si spoločenského postavenia a výskumného potenciálu.

Einstein mal extrémne vysoké spoločenské postavenie: vďaka svojmu zoznamu úspechov; vďaka tomu, *ako* ich dosiahol; a pretože je to fyzik, ktorého meno si pamätá dokonca aj široká verejnosť, ktorý priniesol slávu vede samotnej.

A my máme sklon miešať si slávu s inými vlastnosťami, a máme sklon pripisovať správanie ľudí ich vlastnostiam a nie situáciám.

Takže máme sklon myslieť si, že Einstein, dokonca ešte skôr než bol slávny, už mal vrodený sklon byť Einsteinom – potenciál *rovnako vzácny ako jeho sláva*, a *rovnako čarovný ako jeho skutky*. Takže ak tvrdíte, že máte *potenciál* urobiť to, čo urobil Einstein, *je to to isté ako tvrdiť, že máte nárok na Einsteinovo postavenie*, čo je vyťahovanie sa vysoko nad spoločenské postavenie, ktoré vám pridelil váš kmeň.

Nevyjadrujem to dobre, ale snažím rozseknúť pomýlenú myšlienku: Einstein patrí do oddeleného magistéria, do posvätného magistéria. Toto posvätné magistérium je oddelené od svetského magistéria; nemôžete sa rozhodnúť stať Einsteinom len tak, ako sa rozhodnete stať profesorom alebo riaditeľom. Iba bytosti s nadprirodzeným potenciálom môžu vstúpiť do tohto magistéria – a aj vtedy je to iba naplnenie osudu, ktorý už mali určený. Ak teda hovoríte, že chcete prekonať Einsteina, tvrdíte tým, že *už ste* súčasťou tohto posvätného magistéria – tvrdíte, že máte rovnakú svätožiaru osudu, s akou sa narodil Einstein, ako keď sa niekto narodí do kráľovskej rodiny...

„Ale Eliezer,“ poviete, „iste nemôže *každý* byť Einstein.“

Chcete tým povedať, nie každý môže robiť veci *lepšie* než Einstein.

„Ehm... hej, to som tým myslel.“

No... v modernom svete máte asi pravdu. Asi by ste si *mali* pamätať, že som transhumanista, takže sa pozerám na ľudí a myslím si: „Viete, je to na figu, že nemá každý potenciál robiť veci lepšie než Einstein, a tento problém vyzerá riešiteľne.“ Zafarbuje to človeku úsudok.

Ale v modernom svete, áno, nemá každý potenciál byť Einsteinom.

Ale... ako to povedať...

Je také slovné spojenie, ktoré som raz počul, nepamätám si kde: „Iba ďalší židovský génius.“ Nejaký básnik alebo spisovateľ alebo filozof alebo niekto iný, vynikajúci v mladom veku, urobil niečo, čo vo veľkej schéme vecí nebolo ohromne dôležité, ani príliš vplyvné, takže ho niekto zotrel slovami: „Iba ďalší židovský génius“.

Keby sa Einstein rozhodol zaútočiť na svoj problém z nesprávnej strany – keby si nebol vybral na riešenie dostatočne dôležitý problém – keby nebol vydržal celé roky – keby bol nesprávne odbočil na niektorej z mnohých križovatiek – alebo keby niekto iný vyriešil tento problém pred ním – potom by drahý Albert skončil ako iba ďalší židovský génius.

Géniovia sú zriedkaví, ale nie až *takí* zriedkaví. Nie je až také neuveriteľné tvrdiť, že máte ten druh intelektu, ktorý vás môže dostať do situácie „iba ďalší židovský génius“ alebo „iba ďalší skvelý mozog, ktorý vo svojom živote nikdy neurobil nič zaujímavé“. Spoločenské postavenie spojené s týmto nie je dosť vysoké na to, aby bolo posvätné, takže by to malo vyzerať ako bežne vyhodnotiteľné tvrdenie.

Čo odlišuje ľudí ako sú títo od Einsteina, nie je podľa mňa žiaden vrodený nedostatok skvelosti. Sú to veci ako „nemám zaujímavý problém“ - alebo, aby som správne umiestnil vinu, „nevybral som si zaujímavý problém“. Na tomto je veľmi ľahké zlyhať, pretože máme problém s uloženými myšlienkami: Povedzte ľuďom, aby si vybrali dôležitý problém, a oni vyberú prvý záznam v uloženej pamäti pre „dôležitý problém“, ktorý im vyskočí v hlave, ako „globálne otepľovanie“ alebo „teória strún“.

Skutočne dôležité problémy sú často tie, nad ktorými ani neuvažujete, pretože sa zdajú neriešiteľné, alebo, ehm, *naozaj zložité*, alebo najhoršie zo všetkého, *nie je jasné, ako ich riešiť*. Keby ste na nich pracovali celé roky, možno by vám nepripadali až také nemožné... ale toto je nezvyčajný postreh navyše; naivný realizmus vám povie, že riešiteľné problémy vyzerajú riešiteľne, a neriešiteľne vyzerajúce problémy sú neriešiteľné.

Potom musíte prísť s novým a *hodnotným* smerom útoku. Mnohí ľudia, ktorí nemajú alergiu voči novinkám, zájdu príliš ďaleko opačným smerom, a padnú do afektívnej špirály smrti.

A potom musíte búchať svojou hlavou o tento problém celé roky, a nenechať sa rozptyľovať pokušeniami jednoduchšieho života. „Život je to, čo sa deje, kým si robíme iné plány“ hovorí porekadlo, takže ak chcete naplniť svoje iné plány, musíte byť často pripravení odmietnuť život.

Spoločnosť tiež nie je nastavená tak, aby vás pri práci podporovala.

Pointa je, že problémom nie je v tom, že potrebujete mať svätožiaru osudu, a že vám táto svätožiara osudu chýba. Keby ste stretli Alberta predtým než uverejnil svoje články, nevideli by ste nad ním žiadnu svätožiaru osudu zodpovedajúcu jeho budúcemu vysokému spoločenskému postaveniu. Vyzeral by ako iba ďalší židovský génius.

To nie je preto, že by znamenie kráľovského rodu bolo *ukryté*, ale preto, že jednoducho *žiadne nie je*. *Netreba* ho. *Neexistuje* oddelené magistérium pre ľudí, ktorí robia dôležité veci.

Hovorím to preto, lebo chcem vo svojom živote robiť dôležité veci, a mám naozaj dôležitý problém, a mám smer útoku, a búcham si oň hlavu už celé roky, a podarilo sa mi vybudovať si systém, ktorý ma podporuje; a často stretávam ľudí, ktorí mi takým či onakým spôsobom povedia: „Hej? Chcem vidieť tvoju svätožiaru osudu, kamoš.“

Čo na mňa u Juliana Barboura zapôsobilo, bola vlastnosť, ktorú podľa mňa nikto nevie správne predstierať, pokiaľ ju *nemá*: Barbour vyzeral, že má Einsteina *prekuknutého* – hovoril o Einsteinovi akoby všetko, čo Einstein urobil, bolo dokonale zrozumiteľné a obyčajné.

Hoci ešte aj keď som si to uvedomil, stále ma šokovalo, keď Barbour povedal niečo v štýle: „Tak tuto sa Einsteinovi nepodarilo aplikovať svoje vlastné metódy, čím mu unikol dôležitý poznatok...“ Ale ten šok rýchlo prešiel, pretože som poznal Zákon: *Neexistujú bohovia, neexistuje mágia, a dávni hrdinovia sú iba míľniky, ktoré si odškrtávame v spätnom zrkadle.*

Toto *prekuknutie* je niečo, čo človek musí *dosiahnuť*, je to poznatok, ktorý musí objaviť. Nemôžete prekuknúť Einsteina tým, že iba poviete: „Einstein je obyčajný!“, ak vám jeho dielo stále pripadá ako zázrak. To by bolo ako vyhlásiť: „Vedomie sa musí redukovať na neuróny!“ bez nejakej predstavy, ako na to. Je to síce pravda, ale nerieši to daný problém.

Nebudem vám hovoriť, že Einstein bol bežný chlap, ktorého médiá zveličili, alebo že to bol v zásade rovnaký pako ako každý iný. To by som zašiel *príliš* ďaleko. Aby človek prešiel touto cestou, musí nadobudnúť vlastnosti, ktoré niektorí považujú za... neprirodzené. Ja mám mimoriadnu radosť z robenia vecí, ktoré ľudia nazývajú „mimo ľudských možností“, lebo to ukazuje, že rastiem.

Napriek tomu, zázračné schopnosti nadobúdate nie tak, že sa s nimi narodíte, ale že uvidíte, s náhlym šokom, že *v skutočnosti sú* celkom normálne.

Toto je všeobecný princíp života.

*→

256. Skupinový projekt

„Tak dobre ako Einstein?“ povedal Jeffreyssai neveriacky. „*Iba* tak dobre ako Einstein? Albert Einstein bol veľkým vedcom svojej doby, ale to bola jeho doba, nie dnešná! Einstein nepoznal bayesovské metódy; žil skôr než boli objavené kognitívne skreslenia; nerozumel z vedeckého hľadiska svojim vlastným myšlienkovým postupom. Einstein rozprával nezmysly o neosobnom Bohu – čo vám ukazuje, ako málo chápal rytmus rozumu, keď ho vypínal mimo svojho odboru! Bol príliš zamestnaný divadlom odmietania kvantovej mechaniky svojej doby, než aby ju naozaj *opravil*. A hoci uznávam, že Einstein uvažoval rozumne vo veci všeobecnej relativity – nerátajúc tú vec s kozmologickou konštantou – trvalo mu to desať rokov. Príliš pomaly!“

„*Príliš pomaly?*“ zopakoval Taji neveriacky.

„Príliš pomaly! Keby bol Einstein teraz v tejto triede, a nie na Zemi v mínus prvom storočí, dostal by trstenicou po prstoch! *Vy sa nebudete snažiť robiť to tak dobre ako Einstein! Pokúsite sa to robiť LEPŠIE než Einstein, inak sa ani nenamáhajte!*“

Jeffreyssai potriasol hlavou. „No, už som vám dal dosť nápovedy. Je načase vyskúšať vaše schopnosti. Ja viem, že ostatní *beisutsukai* moje skupinové projekty príliš neschvaľujú...“ Jeffreyssai urobil významnú pauzu.

Brennan si v duchu vzdychol. Túto vetu počul už mnohokrát predtým, v Bardskej Konšpirácii, v Kompetitívnej Konšpirácii: *Ostatní učitelia si myslia, že moje zadania sú príliš ľahké, mali by ste byť vďační*, a nasleduje nejaká nehorázne zložitá úloha...

„Hovoria,“ povedal Jeffreyssai, „že moje projekty sú príliš ťažké; šialene ťažké; že prechádzajú z oblasti šialenstva do oblasti Sparty; že samotný Laplace by na nich zhorel; obviňujú ma, že sa snažím duševne zničiť svojich študentov...“

A sakra.

„Ale existuje dôvod,“ povedal Jeffreyssai, „prečo mnohí z mojich študentov dosiahli veľké veci; a nemyslím tým vysoký post v Bayesovskej Konšpirácii. Očakával som od nich veľa, a oni sa sami naučili od seba veľa očakávať. Takže...“

Jeffreyssai sa na chvíľu rozhliadol po svojich čoraz nervóznejších študentoch. „Tu je vaša úloha. O kvantovej mechanike a všeobecnej relativite ste už počuli. To je hranica vedy predkov, a teda hranica verejného poznania. Vy piati, pracujúci ako skupina, máte za úlohu vytvoriť správnu teóriu kvantovej gravitácie. Váš časový limit je jeden mesiac.“

„*Čo?*“ povedali Brenna, Taji, Styrlyn a Yin. Hiriwa na nich prekvapene pozrela.

„Ak sa vám to podarí,“ pokračoval Jeffreyssai, „budete povýšení na *beisutsukai* druhého *danu* a šiestej úrovne. Uvidíme, či ste sa naučili rýchlosti. Začínate... *teraz*.“

A Jeffreyssai odkráčal z miestnosti, zabuchnúc za sebou dvere.

→ http://lesswrong.com/lw/qs/einsteins_superpowers/

„To je šialené!“ vykríkol Taji.

Hiriwa zmätene pozrela na Tajiho. „Nepoznáme, aké je riešenie. Ako môžeš vedieť, že je také ťažké?“

„Pretože tento problém bol *známy* ešte za dávnych čias! Dávni vedci na tomto probléme robili omnoho dlhšie než jeden mesiac.“

Hiriwa pokrčila plecami. „Takisto sa stále hádali aj o mnohých svetoch, alebo nie?“

„*Dosť! Na toto nemáme čas!*“

Zvyšní štyria študenti pozreli na Styrlyna, pamätajúc si, že sa o ňom hovorí, že má vysoké postavenie v Kooperatívnej Konšpirácii. Nasledovala krátka chvíľa zvažovania, odhadovania, a potom sa Styrlyn stal ich vodcom.

Styrlyn sa zhlboka nadýchol. „Potrebujeme zoznam postupov. Napíšte všetky uhly pohľadu, ktoré vám napadnú. Samostatne – potrebujeme vaše jednotlivé zložky skôr než ich začneme spájať. O päť minút sa každého z vás opýtam, aký je jeho najlepší nápad. *Nestrácajme myšlienky! Ideme!*“

Brennan schmatol hárok a písatko, priložil hrot k povrchu a potom sa zastavil. Nenapadalo mu nič chytré ohľadom zjednotenia všeobecnej relativity a kvantovej mechaniky...

Zvyšní študenti už písali.

Brennan poklepkal hrotom raz, dvakrát, trikrát. Všeobecná relativita a kvantová mechanika...

Taji odložil prvý hárok bokom a schmatol druhý.

Nakoniec Brennan, pre nedostatok čohokoľvek chytrého, napísal to, čo bolo samozrejmé.

O pár minút neskôr, keď Styrlyn ohlásil koniec, to bolo stále všetko, čo mal napísané.

„Dobre,“ povdal Styrlyn, „váš najlepší nápad. Alebo ten nápad, o ktorom najviac chcete, aby sme ho my ostatní vzali do úvahy v druhom kole. Taji, ideš!“

Taji sa pozrel na svoje hárky. „Dobre, myslím, že musíme predpokladať, že všetky cesty, ktoré skúšali dávni vedci, boli slepé uličky, lebo inak by to boli našli. A ak sa toto dá urobiť za jeden mesiac, odpoveď musí byť, v istom zmysle, elegantná. Takže žiadne viaceré rozmery. Ak začneme robiť niečo, čo vyzerá, že by sme to mali nazvať ‚teória strún‘, radšej sa zastavme. Azda by sme mohli začal úvahou, ako nechápanie dekoherencie mohlo pomýliť dávnych vedcov pri kvantizovaní gravitácie.“

„Opakom hlúposti je hlúposť,“ povedala Hiriwa. „Predpokladajme, že dávna veda nikdy neexistovala.“

„Zatiaľ ešte nekritizujme!“ povedal Styrlyn. „Hiriwa, tvoj návrh?“

„Zbavme sa nekonečien,“ povedala Hiriwa, „zrušme všetko, čo ich dovoľuje. Nemala by to byť otázka chytrého narábania s integrálmi. *Reprezentácia*, ktorá umožňuje nekonečno, musí byť fakticky nepravdivá.“

„Yin.“

„Vieme bežným rozumom,“ povedala Yin, „že keby sme vystúpili mimo vesmíru, videli by sme čas rozprestený celý pred sebou, skutočnosť by bola ako kryštál. Ale raz som natrafila na náznak, že fyzika je bezčasová ešte v hlbšom zmysle, než je toto.“ Yin neprítomne pozerala, spomínala. „Pred rokmi som našla opustené mesto; myslím, že bolo neobývané celé veky. A za dverami so zlomenými zámkami bolo na jednej stene vyryté: ‚*.ua sai .ei mi vimcu ty bu le mekso*‘.“

Brennan si preložil: *Heuréka! Odstráň t z rovníc.* Napísané v Lojbane, posvätnom jazyku vedy, čo znamenalo, že neznámy autor si myslel, že je to pravda.

„Tá ‚bezčasová fyzika‘, o ktorej sme všetci počuli povesti,“ povedala Yin, „môže byť bezčasová vo veľmi doslovnom zmysle.“

„Môj príspevok,“ povedal Styrlyn. „Kvantová fyzika, ktorú sme sa naučili, je nad spoločnými reprezentáciami pozícii. Zdá sa, že by sme to mali vedieť rozobrať na priestorovo lokálne reprezentácie, v zmysle invariantných previazaní na diaľku. Nájdenie takejto reprezentácie by nám mohlo pomôcť spojiť to so všeobecnou relativitou, ktorej zakrivenosť je lokálna.“

„Veľmi *individualistický* pohľad,“ zamrmlal Taji, „na niekoho z Kooperatívnej Konšpirácie.“

Styrlyn potriasol hlavou. „Potom nás nechápeš. Prvá lekcia, ktorú sa učíme, je že skupiny sa skladajú z ľudí... nie, teraz nemáme čas na politiku. Brennan!“

Brennan pokrčil plecami. „Nič moc, obávam sa, iba samozrejmosti. Zotrvačná hmotnosť/energia sa pri pozorovaniach vždy rovnala gravitačnej hmotnosti/energii, a Einstein ukázal, že sú nevyhnutne to isté. Prečo je potom ‚energia', ktorá je eigenhodnotou kvantového hamiltoniánu, *nevyhnutne* rovnaká ako údaj ‚energia', ktorý sa objavuje v rovniciach všeobecnej relativity? Prečo by sa mal časopriestor zakrivovať v takej istej miere, ako sa malé šípky otáčajú?"

Bola krátka pauza.

Yin si odfrkla. „Toto znie *príliš* samozrejme. Nedošli by na to už dávni vedci?"

„Zabudni, že dávna veda existovala," povedala Hiriwa. „Je tu otázka: potrebujeme odpoveď, či už v dávnych dobách bola alebo nebola známa. Nemôže to byť iba *zhoda okolností*."

Taji pozeral do prázdna. „Možno by sa dalo ukázať, že výnimka z tejto rovnice by porušila nejaký zákon zachovania..."

„To nie je to, čo Brennan naznačoval," prerušila ho Hiriwa. „Nepýtal si dôkaz, že sa musia rovnať kvôli nejakému presvedčivému princípu; pýtal si pohľad, v ktorom tieto dve veci sú jedna a nemožno ich oddeliť ani len pojmovo, ako sa dosiahlo pri zotrvačnej hmote/energii a gravitačnej hmote/energii. Musíme totiž predpokladať, že krása tohto celku vychádza zo základných zákonov, a nie naopak. Súhlasíš s takouto formuláciu?"

„Súhlasím," odpovedal Brennan.

Zavládlo ticho na tridsaťsedem sekúnd, ako sa všetci piati zamysleli nad piatimi návrhmi.

„Mám jeden nápad..."

*→

Medzihra: Technické vysvetlenie technického vysvetlenia

Ako zdôrazňuje Jaynes (1996), vety bayesovskej teórie pravdepodobnosti sú presne to, *matematické vety*, ktoré nevyhnutne vyplývajú z bayesovských axióm.[216] Človek by si mohol naivne myslieť, že ohľadom matematických viet nebude žiadna kontroverzia. Ale kedy použiť tieto vety? Ako použijeme tieto vety pri problémoch v skutočnom svete? Intuitívne vysvetlenie sa snaží vyhnúť kontroverzii, ale *Technické vysvetlenie* ochotne napochoduje medzi víriace čepele vrtuľníka. Bez obalu, rozmýšľanie v *Technickom vysvetlení* nereprezentuje jednohlasný konsenzus spoločenstva bayesovských výskumníkov na celej planéte Zem. Prinajmenšom zatiaľ nie.

Intuitívne vysvetlenie sa sústredilo na poskytnutie pevného chápania základov bayesovstva, *Technické vysvetlenie* buduje na bayesovských základoch tézy o ľudskej rozumnosti a o filozofii vedy. *Technické vysvetlenie technického vysvetlenia* sa volá takto, pretože začalo touto otázkou:

„Aký je rozdiel medzi technickým pozorumením, a slovným porozumením?"

* * *

Ako dieťa som čítal populárne fyzikálne knihy, a zdalo sa mi, že toho viem veľa; myslel som si, že viem, že zvuk sú vlny vzduchu, svetlo sú elektromagnetické vlny, hmota sú vlny komplexných amplitúd pravdepodobnosti. Keď som vyrástol, čítal som *Feynmanove lekcie z fyziky* a objavil som klenot s názvom „vlnová rovnica".[217] A vtedy som si uvedomil, že až dovtedy som nerozumel ani neveril, že „zvuk sú vlny" spôsobom podobným tomu, ako túto vetu myslí a verí fyzik.

Toto je teda rozdiel medzi technickým porozumením a slovným porozumením.

Veríte tomu? Ak áno, mali by ste aplikovať túto vedomosť a povedať: „Ale prečo si nám nedal technické vysvetlenie namiesto slovného vysvetlenia?"

* * *

→ http://lesswrong.com/lw/qt/class_project/

216 Edwin T. Jaynes, *Probability Theory: The Logic of Science*, ed. George Larry Bretthorst (New York: Cambridge University Press, 2003), doi:10.2277/0521592712.

217 Feynman, Leighton, and Sands, *The Feynman Lectures on Physics*.

Predstavte si *hustotu pravdepodobnosti* alebo *masu pravdepodobnosti* – pravdepodobnosť ako hruda hliny, ktorú musíte rozdeliť medzi možné výsledky.

Povedzme, že je malé svetlo, ktoré môže bliknúť na *červeno*, na *modro*, alebo na *zeleno* vždy keď stlačíte tlačidlo. Svetlo pri každom stlačení tlačidla blikne iba raz a iba jednou farbou; tieto možnosti sa navzájom vylučujú. Pokúšate sa predpovedať farbu nasledujúceho bliknutia. Pri každom pokuse máte hrudu hliny, masu pravdepodobnosti, ktorú musíte rozdeliť medzi možnosti: červená, zelená, modrá. Mohli by ste dať štvrtinu svojej hliny na možnosť „zelená“, štvrtinu svojej hliny na možnosť „modrá“, a polovicu svojej hliny na možnosť „červená“ - ako priradenie pravdepodobnosti: zelená: 25 %, modrá: 25 %, a červená: 50 %. Metafora je, že *pravdepodobnosť je obmedzený zdroj*, ktorý treba prideľovať šetrne. Ak si myslíte, že je väčšia šanca, že pri ďalšom pokuse blikne „modrá“, môžete modrej prideliť väčšiu pravdepodobnosť, ale musíte ubrať masu pravdepodobnosti z iných hypotéz – možno ukradnúť trochu hliny z červenej, a pridať ju k modrej. Nikdy nedostanete žiadnu ďalšiu hlinu. Vaše pravdepodobnosti nemôžu mať súčet väčší než 1,0 (100 %). Nemôžete predpovedať šancu 75 %, že uvidíte červenú, a šancu 80 %, že uvidíte modrú.

Prečo by ste chceli byť opatrní so svojou masou pravdepodobnosti, alebo rozdeľovať ju šetrne? Prečo nerozhádzať pravdepodobnosť všade naokolo? Zmeňme metaforu z hliny na peniaze. Môžete staviť dokopy jeden dolár hracích peňazí pri každom stlačení tlačidla. Vedľa stojí experimentátor a vyplatí vám skutočné peniaze, ktorých množstvo závisí od toho, koľko hracích peňazí ste stavili na *vyhrávajúce* svetlo. Nezaujíma nás, ako ste rozdelili zvyšné peniaze medzi prehrávajúce svetlá. Jediná dôležitá vec je, koľko ste stavili na to svetlo, ktoré naozaj vyhralo.

Musíme však opatrne vytvoriť bodovacie pravidlo používané na výplatu víťazov, ak chceme, aby si hráči dávali pri stávkovaní pozor. Predstavte si, že by experimentátor vyplatil hráčovi skutočné peniaze rovnajúce sa hracím peniazom na vyhrávajúcom svetle. Pri tomto pravidle, ak si všimnete, že červená vychádza v 6 prípadoch z 10, vaša najlepšia stratégia je staviť na červenú nie 60 centov, ale celý dolár, a frekvencie zelenej a modrej vás nezaujímajú. Prečo? Povedzme, že aj modrá aj zelená vychádzajú zhruba v 2 prípadoch z 10. A predstavte si, že ste stavili 60 centov na červenú, 20 centov na modrú, a 20 centov na zelenú. Takto v 6 prípadoch z 10 vyhráte po 60 centov, a v 4 prípadoch z 10 vyhráte po 20 centov, čiže priemerná výhra je 44 centov. Pri tomto bodovacom pravidle dáva väčší zmysel umiestniť celý dolár na červenú, a vyhrať celý dolár v 6 prípadoch z 10. V 4 prípadoch z 10 nevyhráte nič. Vaša priemerná výhra bude 60 centov.

Ak si napíšeme funkciu výplaty, bolo by to Výplata = P(víťaz), kde P(víťaz) je množstvo hracích peňazí, ktoré ste stavili na vyhrávajúcu farbu v tomto kole. Ak si napíšeme funkciu očakávanej výplaty podľa tohto pravidla, bolo by to

Očakávanie[Výplata] = (Suma[P(farba) × F(farba)] pre každú farbu).

P(farba) je množstvo hracích peňazí, ktoré ste stavili na danú farbu, a F(farba) je frekvencia, s akou táto farba vyhráva.

Predstavte si, že skutočná frekvencia svetiel je: modré: 30 %, zelené: 20 %, a červené: 50 %. A predstavte si, že v každom kole stavím: modré: 0,40 $, zelené: 0,50 $, a červené: 0,10 $. V 30 % prípadov by som dostal 0,40 $, v 20 % prípadov 0,50 $, a v 50 % prípadov 0,10 $, čiže priemerná výplata by bola 0,12 $ + 0,10 $ + 0,05 $ alebo 0,27 $.

Čiže:

P(farba) = hracie peniaze priradené k tejto farbe

F(farba) = frekvencia, s ktorou táto farba vyhráva

Výplata = P(víťaz) = množstvo hracích peňazí pridelených k vyhrávajúcej farbe

Skutočné frekvencie vyhrávania:

Modrá: 30 %

Zelená: 20 %

Červená 50 %

Z dlhodobého hľadiska červená vyhráva v 50 % prípadov, zelená vyhráva v 20 % prípadov, a modrá vyhráva v 30 % prípadov. Naša *priemerná* výplata v každom kole je teda 50 % výplaty, ak vyhrá červená, plus 20 % výplaty, ak vyhrá zelená, plus 30 % výplaty, ak vyhrá modrá.

Výplata je funkcia vyhrávajúcej farby a systému stávkovania. Chceme vypočítať *priemernú* výplatu pri danom systéme stávkovania a *frekvenciách* ako vyhrávajú jednotlivé farby. Matematický pojem pre tento druh výpočtu, keď vezmeme funkciu pre každý prípad a dáme jej váhu podľa frekvencie daného prípadu, je *očakávanie*. Teda, aby sme vypočítali našu *očakávanú výplatu*, spočítame:

Očakávanie[Výplata] = Suma[P(farba) × F(farba)] pre každú farbu

= P(modrá) × F(modrá) + P(zelená) × F(zelená) + P(červená) × F(červená)

= 0,40 $ × 30 % + 0,50 $ × 20 % + 0,10 $ × 50 %

= 0,12 $ + 0,10 $ + 0,05 $

= 0,27 $.

S týmto systémom stávkovania vyhrám v priemere 27 centov každé kolo.

Svoje hracie peniaze som umiestnil hrubo svojvoľne, a vzniká otázka: Môžem zvýšiť svoju očakávanú výplatu múdrejším umiestnením mojich hracích peňazí? *Pri danom pravidle bodovania* svoju očakávanú výhru maximalizujem, ak umiestnim *celý* svoj dolár na červenú. Hoci mám *očakávanú* výplatu päťdesiat centov za kolo, svetlo by mohlo *v skutočnosti* zablikať: zelená, modrá, modrá, zelená, zelená, a ja by som dostal *skutočnú* výplatu nula. Ale šanca, že svetlo v piatich nasledujúcich kolách nebude ani raz červené je približne 3 %. Porovnajte s kartovou hrou s červenými a modrými kartami v kapitole Zákonitá neistota.

Správne bodovacie pravidlo (ďalší štandardný matematický pojem) je také pravidlo bodovania stávok, že svoju očakávanú výplatu maximalizujete tým, že stavíte hracie peniaze, ktoré sa presne rovnajú šanci, že zabliká daná farba. Chceme také bodovacie pravidlo, že ak svetlo naozaj bliká s frekvenciou: modrá: 30 %, zelená: 20 %, červená 50 %, potom svoju priemernú výplatu môžete maximalizovať *iba* tak, že stavíte 30 centov na modrú, 20 centov na zelenú, a 50 centov na červenú. Správne bodovacie pravidlo je také, ktoré donúti vašu optimálnu stávku presne nahlásiť váš odhad pravdepodobnosti. (To sa niekedy nazýva aj *striktne správne bodovacie pravidlo*.) Ako sme videli, nie každé bodovacie pravidlo má túto vlastnosť; a ak si vymyslíte náhodné dôveryhodne vyzerajúce pravidlo, pravdepodobne túto vlastnosť mať *nebude*.

Jedno pravidlo pri tejto vlastnosti je platiť dolár mínus druhá mocnina chyby stávky, nie samotnú stávku – ak ste na vyhrávajúce svetlo stavili 30 centov, vaša chyba je 70 centov, druhá mocnina chyby je 49 centov ($0,7^2 = 0,49$), a dolár mínus druhá mocnina chyby je 51 centov.[218] (Predpokladáme, že vaše hracie peniaze sú denominované v odmocninách centov, takže ich odmocnina je platná finančná hodnota.)

My *nebudeme* používať pravidlo druhej mocniny chyby. Bežní štatistici používajú druhú mocninu chyby všetkého, čo vidia, ale nie bayesovskí štatistici.

My si pridáme novú požiadavku: budeme chcieť nielen správne bodovacie pravidlo, ale aby naše správne bodovacie pravidlo dávalo rovnaký výsledok bez ohľadu na to, či ho uplatíme na jednotlivé kolá alebo ich kombinácie. To je to, čo robia Bayesovci namiesto používania druhej mocniny chyby niečoho; požadujeme invarianty.

Predstavme si, že stlačím tlačidlo dvakrát po sebe. Máme deväť možných výsledkov: zelená-zelená, zelená-modrá, zelená-červená, modrá-zelená, modrá-modrá, modrá-červená, červená-zelená, červená-modrá, červená-červená. Predstavte si, že vyhrá zelená a potom vyhrá modrá. Experimentátor by priradil prvé skóre podľa nášho odhadu pravdepodobnosti $p(\text{zelená}_1)$ a druhé skóre podľa $p(\text{modrá}_2 \mid \text{zelená}_1)$.[219] Urobili by sme dve predpovede a dostali dve bodovania. Naša prvá predpoveď by bola pravdepodobnosť,

218 Čitatelia so znalosťou derivácie si môžu overiť, že v jednoduchšom prípade, kde svetlo má iba dve farby, kde na prvú farbu stavíme p, a frekvencia prvej farby je f, očakávaná výplata je $f \times (1 - (1 - p)^2) + (1 - f) \times (1 - p^2)$, kde p je premenná a f je konštanta, má globálne maximum v bode p = f.

219 Nepamätáte sa, ako sa číta p(A | B)? Pozrite si Intuitívne vysvetlenie bayesovského uvažovania.

ktorú sme priradili farbe, ktoré vyhrala v prvom kole, zelenej. Naša druhá predpoveď by bola pravdepodobnosť, že modrá vyhrá v druhom kole, *za predpokladu*, že v prvom kole vyhrala zelená. Prečo potrebujeme písať p(modrá$_2$ | zelená$_1$) namiesto iba p(modrá$_2$)? Pretože by ste mohli mať o blikajúcom svetle hypotézu, ktorá hovorí „modrá nikdy nejde po zelenej“ alebo „modrá ide vždy po zelenej“ alebo „modrá ide po zelenej s pravdepodobnosťou 70 %“. Ak je to tak, potom po videní zelenej v prvom kole môžete chcieť zmeniť svoju predpoveď – zmeniť svoju stávku – pre druhé kolo. Vždy môžete meniť svoje predpovede až do chvíle, keď experimentátor stlačí dané tlačidlo, použijúc každý kúsok informácie, ktorý máte; ale keď už svetlo zablikalo, je príliš neskoro meniť stávku.

Predstavme si, že skutočný výsledok je zelená$_1$ a po nej modrá$_2$. Považujeme tento invariant: Musím dostať rovnaké výsledné skóre bez ohľadu na to, či:

- Dostávam body dvakrát, prvýkrát za svoju predpoveď p(zelená$_1$) a druhýkrát za svoju predpoveď p(modrá$_2$ | zelená$_1$).
- Dostávam body raz, za spoločnú predpoveď p(modrá$_2$ a zelená$_1$).

Predstavme si, že priradím pravdepodobnosť 60 % pre zelená$_1$, a potom blikne zelené svetlo. Teraz musím dodať pravdepodobnosti pre farby v druhom kole. Odhadnem pravdepodobnosť modrá$_2$ a priradím jej 25 % svojej hmotnosti pravdepodobnosti. A hľa, v druhom kole svetlo blikne na modro. Takže v prvom kole bola moja stávka na vyhrávajúcu farbu 60 %, a v druhom kole bola moja stávka na vyhrávajúcu farbu 25 %. Ale takisto som si mohol na začiatku experimentu a po priradení p(zelená$_1$) predstaviť, že svetlo prvýkrát blikne na zeleno, predstaviť si, že aktualizujem svoje teórie na základe tejto informácie, a potom povedať, akú dôveru by som dal tomu, že v ďalšom kole bude modrá, ak v prvom kole bola zelená. To jest, vytvoriť pravdepodobnosti p(zelená$_1$) a p(modrá$_2$ | zelená$_1$). Vynásobením týchto dvoch pravdepodobností dostaneme spoločnú pravdepodobnosť, p(zelená$_1$ a modrá$_2$) = 15 %.

Dvojitý pokus má deväť možných výsledkov. Ak vytvorím deväť pravdepodobností pre p(zelená$_1$ a zelená$_2$), p(zelená$_1$ a modrá$_2$) … p(červená$_1$ a modrá$_2$), p(červená$_1$ a červená$_2$), hmotnosť pravdepodobnosti nesmie dať súčet väčší než 1,0. Robím predpovede deviatich navzájom sa vylučujúcich možností v tomto „dvojitom experimente“.

Budeme požadovať bodovacie pravidlo (a možno to zďaleka nebude vyzerať ako niečo, čo by by bežný bookmaker niekedy použil) také, aby sa moje skóre nezmenilo podľa toho, či tento dvojitý výsledok považujeme za dve predpovede alebo za jednu predpoveď. Môžem vnímať postupnosť dvoch výsledkov ako jeden experiment „stlač tlačidlo dvakrát“ a byť hodnotený podľa mojej predpovede p(modrá$_2$ a zelená$_1$) = 15 %. Alebo môžem byť hodnotený raz za moju prvú predpoveď p(zelená$_1$) = 60 %, a potom zase za moju predpoveď p(modrá$_2$ | zelená$_1$) = 25 %. V oboch prípadoch požadujeme rovnaké *celkové* skóre, aby nezáležalo na tom, ako si rozdelíme experimenty a predpovede – *celkové* skóre bude vždy rovnaké. Toto je náš invariant.

Práve sme žiadali:

Skóre(p(zelená$_1$ a modrá$_2$)) = Skóre(p(zelená$_1$)) + Skóre(p(modrá$_2$ | zelená$_1$))

A už vieme:

$$p(\text{zelená}_1 \text{ a modrá}_2) = p(\text{zelená}_1) \times p(\text{modrá}_2 \mid \text{zelená}_1)$$

Jediné možné bodovacie pravidlo je:

$$\text{Skóre}(p) = \log(p)$$

Nové bodovacie pravidlo je, že vaše skóre je *logaritmom* pravdepodobnosti, ktorú ste priradili víťazovi.

Základ logaritmu je ľubovoľný – či použijeme logaritmy pri základe desať alebo logaritmy pri základe dva, bodovacie pravidlo má žiadaný invariant. Ale musíme si vybrať nejaký konkrétny základ.

Matematik by si vybral základ *e*; inžinier by si vybral základ desať; informatik by si vybral základ dva. Ak použijeme základ desať, môžeme prepočítať na *decibely*, ako v Intuitívnom vysvetlení; ale niekedy sa ľahšie narába s bitmi.

Logaritmické bodovacie pravidlo je správne – má svoje očakávané maximum, keď povieme svoje presné očakávania; odmeňuje úprimnosť. Ak si myslíme, že modré svetlo má pravdepodobnosť bliknutia 60 %, a počítame svoju očakávanú výplatu pre rôzne schémy stávok, zistíme, že môžeme maximalizovať svoju očakávanú výplatu tým, že povieme experimentátorovi „60 %“. (Čitatelia znalí derivácií si to môžu skontrolovať.) Toto bodovacie pravidlo zároveň dáva invariantný súčet, bez ohľadu na to, či stlačenie tlačidla dvakrát považujeme za „jeden experiment“ alebo „dva experimenty“. Výplaty sú však teraz všetky *záporné*, pretože berieme logaritmus pravdepodobnosti, a pravdepodobnosť je medzi 0 a 1. Logaritmus pri základe desať z 0.1 je -1; logaritmus pri základe desať z 0,01 je -2. To je okej. Zmierili sme sa s tým, že sa toto bodovacie pravidlo nemusí podobať na nič, čo by hocijaký skutočný bookmaker niekedy použil. Ak chcete, môžete si predstaviť, že experimentátor má hromadu peňazí, a na konci experimentu vám pridelí nejakú množstvo mínus vaše veľké záporné skóre. (Ehm, množstvo plus vaše záporné skóre.) Možno má experimenátor sto dolárov, a na konci stého kola ste dokopy nazbierali skóre -48, takže dostanete 52 dolárov.

Skóre -48 pri akom základe? Môžeme túto nejasnosť v skóre odstrániť určením jednotiek. Desať decibelov rovná sa činiteľu 10; mínus desať decibelov rovná sa činiteľu 1/10. Priradiť skutočnému výsledku pravdepodobnosť 0,01 dáva skóre -20 decibelov. Pravdepodobnosť 0,03 dáva skóre -15 decibelov. Niekedy môžeme použiť bity: 1 bit je činiteľ 2, -1 bit je činiteľ ½. Pravdepodobnosť 0,25 dáva skóre -2 bity; pravdepodobnosť 0,03 dáva skóre okolo -5 bitov.

Ak dôjdete k pravdepodobnostnému odhadu P pre každú farbu, čiže p(červená), p(modrá), p(zelená), potom vaše *očakávané skóre* je:

Skóre = log(p)

Očakávanie[Skóre] = Suma[p × log(p)] pre všetky výsledky p

Predstavte si, že máte pravdepodobnosti: červená: 25 %, modrá: 50 %, zelená: 25 %. Rozmýšľajme chvíľu pri základe 2, nech sú veci jednoduchšie. Vaše očakávané skóre je:

červená: skóre -2 bity, blikne v 25 % prípadov,

modrá: skóre -1 bit, blikne v 50 % prípadov,

zelená: skóre -2 bity, blikne v 25 % prípadov,

očakávané skóre: -1,5 bitu.

* * *

Porovnajme naše bayesovské bodovacie pravidlo s bežným alebo hovorovým spôsobom vyjadrovania stupňov istoty, kde niekto môže len tak povedať: „Som si na 98 % istý, že repkový olej obsahuje viac omega-3 tukov než olivový olej.“ Čo tým naozaj myslí, je, že sa cíti na 98 % istý – že existuje niečo ako malá stupnica merajúca silu emócie istoty, a táto stupnica je zaplnená na 98 %. A táto emocionálna stupnica by pravdepodobne nebola plná na presne 98 %, keby sme mali nejaký spôsob, ako ju merať. Slová „98 %“ sú iba hovorovým spôsobom vyjadrenia: „Som si takmer ale nie celkom istý.“ Neznamená to, že by ste dostali najvyššiu očakávanú výplatu stavením presne 98 hracích peňazí na tento výsledok. Mali by ste priradiť *kalibrovanú istotu* 98 % iba vtedy, keď máte dostatočnú dôveru v to, že by ste mohli odpovedať na sto podobných otázok s rovnakou zložitosťou, jednu za druhou, každú nezávisle na ostatných, a pomýliť sa v priemere asi dvakrát. Budeme si zaznamenávať, ako často máte pravdu, a keď sa časom ukáže, že keď poviete „na 90 % istý“, máte pravdu v 7 prípadoch z 10, potom povieme, že ste *zle kalibrovaný*.

Ak poviete „pravdepodobnosť 98 %“ tisíckrát, a ste prekvapení iba päťkrát, stále do vás budeme hučať, že ste zle kalibrovaní. Priraďujete príliš veľkú masu pravdepodobnosti možnosti, že sa mýlite.

Mali by ste povedať „pravdepodobnosť 99,5 %“, aby ste maximalizovali svoje skóre. Bodovacie pravidlo odmeňuje *presnú* kalibráciu, nepovzbudzuje ani pokoru ani drzosť.

V tomto bode môže niektorým čitateľom napadnúť, že existuje zrejmý spôsob, ako dosiahnuť dokonalú kalibráciu – skrátka si pre každej odpovedi typu „áno alebo nie“ hoďte mincou, a priraďte svojej odpovedi dôveryhodnosť 50 %. Poviete 50 % a v polovici prípadov budete mať pravdu. Nie je toto dokonalá kalibrácia? Áno. Lenže kalibrácia je iba jedna zložka nášho bayesovského skóre; tou druhou zložkou je *rozlišovanie*.

Predstavte si, že vám dám desať otázok typu „áno alebo nie“. O danej téme neviete absolútne nič, takže pri každej odpovedi rozdelíte svoju masu pravdepodobnosti fifty-fifty medzi „Áno“ a „Nie“. Blahoželám, ste dokonale kalibrovaný – odpovede, pri ktorých ste uviedli „pravdepodobnosť 50 %“, boli správne presne v polovici prípadov. To je pravda bez ohľadu na postupnosť správnych odpovedí alebo koľko odpovedí bolo Áno. V desiatich experimentoch ste dvadsaťkrát povedali „50 %“ - povedali ste „50 %“ na Áno_1, Nie_1, Áno_2, Nie_2, Áno_3, Nie_3... V desiatich prípadoch z toho bola odpoveď správna, konkrétne: Áno_1, Nie_2, Nie_3... A v desiatich prípadoch z toho bola odpoveď nesprávna: Nie_1, Áno_2, Áno_3...

Teraz vám dám svoje vlastné odpovede, do ktorých vložím viac úsilia, pokúsim sa rozlíšiť, či je správna odpoveď Áno alebo Nie. Priradím každej zo svojich odpovedí spoľahlivosť 90 % a moja odpoved je nesprávna dvakrát. Som kalibrovaný horšie než vy. Povedal som v desiatich prípadoch „90 %“ a mýlil som sa dvakrát. Keď ma niekto bude nabudúce počúvať, možno si v hlave preloží „90 %“ na 80 %, vediac, že keď som si na 90 % istý, mám pravdu zhruba v 80 % prípadov. Ale pravdepodobnosť, ktorú ste vy priradili konečnému výsledku, je 1/2 na desiatu, čo je 0,001 alebo 1/1024. Pravdepodobnosť, ktorú som ja priradil konečnému výsledku je 90 % na ôsmu krát 10 % na druhú, $0{,}9^8 \times 0{,}1^2$, čo vychádza na 0,004 alebo 0,4 %. Vaša kalibrácia je dokonalá a moja nie, ale moje lepšie *rozlišovanie* medzi správnymi a nesprávnymi odpoveďami to viac než vynahradí. Moje výsledné skóre je vyššie – priradil som väčšiu celkovú pravdepodobnosť konečnému výsledku celého experimentu. Keby som bol menej prehnane sebavedomý a lepšie kalibrovaný, pravdepodobnosť, ktorú by som prisúdil konečnému výsledku, by bola $0{,}8^8 \times 0{,}2^2$, čo dáva 0,006 alebo 0,6 %.

Je možné dopadnúť ešte lepšie? Iste. Mohli by ste odpovedať každú jednu otázku správne, a priradiť pravdepodobnost 99 % každej zo svojich odpovedí. Potom by pravdepodobnosť, ktorú ste priradili celkovému výsledku experimentu bola $0{,}99^{10} \sim 90$ %.

Vaše *skóre* by bolo log(90 %), -0,45 decibelu alebo -0,15 bitu. Potrebujeme počítať logaritmus, aby som nemal motív podvádzať, keď sa snažím maximalizovať moje *očakávané skóre*, Suma[p × log(p)]. Bez logaritmu by som maximalizoval svoje očakávané skóre tým, že by som priradil celú svoju masu pravdepodobnosti na ten najpravdepodobnejší výsledok. Ďalej, bez pravidla algoritmu by bolo moje celkové skóre rôzne podľa toho, či by sme počítali viaceré kolá ako viaceré experimenty alebo ako jeden experiment.

Jednoduchá transformácia dokáže napraviť zlú kalibráciu znížením rozlíšenia. Ak máte vo zvyku hovoriť „milión k jednej“ na 90 správnych a 10 nesprávnych odpovedí na každých sto otázok, môžeme zdokonaliť vašu kalibráciu tým, že nahradíme slová „milión k jednej“ slovami „deväť k jednej“. Neexistuje však žiaden ľahký spôsob, ako zvýšiť (úspešné) rozlišovanie. Ak máte vo zvyku hovoriť „deväť k jednej“ na 90 správnych odpovedí na každých sto otázok, ľahko môžem zvýšiť vaše *udávané* rozlišovanie tým, že nahradím slová „deväť k jednej“ slovami „milión k jednej“. Ale žiadna jednoduchá transformácia nezvýši vaše *skutočné* rozlišovanie tak, aby vaša odpoveď obsahovala 95 správnych odpovedí a 5 nesprávnych. Yates a kol.:[220] „Zatiaľ čo dobrú kalibráciu možno často dosiahnuť jednoduchými matematickými transformáciami (napríklad pridaním konštanty ku každému pravdepodobnostnému úsudku), dobré rozlíšenie si vyžaduje prístup k solídnym, prediktívnym indíciám,

220 J. Frank Yates et al., „Probability Judgment Across Cultures,“ in Gilovich, Griffin, and Kahneman, *Heuristics and Biases*, 271–291.

a schopnosť využívať tieto indície, čo je v praktických situáciách v skutočnom živote ťažké nájsť.“ Ak vám chýba schopnosť rozlíšiť pravdu od nepravdy, môžete dosiahnuť dokonalú kalibráciu tým, že priznáte svoju nevedomosť; ale samotné priznanie nevedomosti nerozlíši pravdu od nepravdy.

Takto sa teda zbavíme ďalšieho falošného stereotypu rozumnosti, že rozumnosť spočíva v tom, že sme pokorní a skromní a priznávame bezmocnosť zoči-voči neznámemu. Toto je iba podvodníkove riešenie, priradenie pravdepodobnosti 50 % všetkým otázkam typu „áno alebo nie“. Naše bodovacie pravidlo vás povzbudzuje, aby ste to robili lepšie, ak môžete. Ak neviete, priznajte svoju nevedomosť; ak si dôverujete, priznajte svoju sebadôveru. Potrestáme vás, ak sa budete sebavedome mýliť, ale takisto vás odmeníme, ak budete mať sebavedome pravdu. Toto je cnosť správneho hodnotiaceho pravidla.

* * *

Predstavte si, že hodím mincou dvadsaťkrát. Ak verím, že minca je vyvážená, najlepšia predpoveď, akú môžem urobiť, je predpovedať pri každom hode rovnakú šancu, že padne hlava alebo znak. Ak verím, že minca je vyvážená, priradím rovnakú pravdepodobnosť každej možnej postupnosti dvadsiatich hodení mincou. Existuje asi milión (1 048 576) možných postupností dvadsiatich hodov mincou, a ja mám na hranie iba 1,0 masy pravdepodobnosti. Preto priradím každej *jednej* možnej postupnosti pravdepodobnosť $1/2^{20}$ – šancu asi milión k jednej; -20 bitov alebo -60 decibelov.

Urobil som experimentálnu predpoveď a dostal som skóre -60 decibelov! Nefalzifikuje to danú hypotézu? Intuitívne, nie. Nerobíme to, že by sme dvadsaťkrát hodili mincou, uvideli náhodný výsledok, potom sa vrátili do minulosti a povedali: hľa, šanca, že sa toto stane, je milión k jednej. Ale šanca uvidieť opäť presne tú istú postupnosť *je* milión k jednej, ako by som zistil, keby som naivne predpovedal, že *ďalšia* postupnosť dvadsiatich hodov mincov dá celkom rovnaký výsledok. Je v poriadku, ak máme teórie, ktoré priraďujú drobné pravdepodobnosti výsledkom, pokiaľ to žiadna iná teória nedokáže lepšie. Ale keby niekto použil alternatívnu hypotézu na zapísanie tej istej postupnosti do zalepenej obálky vopred, a priradil by pravdepodobnosť 99 %, pochyboval by som o vyváženosti mince. Za predpokladu, že by zalepila iba *jednu* obálku, a nie milión.

Toto nám hovorí, *čo* by sme mali zdravým rozumom odpovedať, ale nehovorí nám to, *ako* táto odpoveď zdravého rozumu vzniká z matematiky. Aby sme videli, *prečo* je táto odpoveď zdravého rozumu správna, potrebujeme spojiť všetko, čo sme si zatiaľ povedali o štruktúra bayesovskej zmeny názoru. Keď to budeme mať, budeme mať technické porozumenie rozdielu medzi slovným porozumením a technickým porozumením.

* * *

Predstavte si pokus, ktorý vytvorí celé číslo od nuly do 99. Napríklad, experiment môže byť počítadlo častíc, ktoré povie, koľko častíc ním prešlo za minútu. Alebo experiment môže byť návšteva supermarketu v stredu, zistenie ceny 30-dekového vrecka drvených orechov, a zapísanie posledných dvoch číslic ceny.

Testujeme niekoľko rôznych hypotéz, ktoré sa snažia predpovedať experimentálny výsledok. Každá hypotéza dáva distribúciu pravdepodobnosti pre všetky možné výsledky; v tomto prípade pre celé čísla od nuly do 99. Tieto možnosti sa navzájom vylučujú, takže masa pravdepodobnosti v distribúcii musí dávať súčet 1,0 (alebo menej); nemôžeme predpovedať pravdepodobnosť 90 %, že uvidíme 42, a zároveň pravdepodobnosť 90 %, že uvidíme 43.

Predstavme si, že existuje presná hypotéza, ktorá predpovedá šancu 90 %, že uvidíme výsledok 51. (Napríklad hypotéza je, že v supermarkete zvyčajne nastavujú ceny orechov v tvare „X dolárov a 51 centov“.) Táto presná teória stavila 90 % svojej masy pravdepodobnosti na výsledok 51. To necháva 10 % masy pravdepodobnosti na rozdelenie medzi 99 zvyšných výsledkov – všetky čísla od nuly do 99 *okrem* 51. Táto teória ďalej nič neupresňuje, takže rozdelíme zvyšných 10 % masy pravdepodobnosti rovnomerne medzi 99 možností, priradiac pravdepodobnosť 1/990 každému výsledku okrem 51. Pre jednoduchosť zapisovania zaokrúhlime 1/990 na 0,1 %.

Táto distribúcia pravdepodobnosti je analogická *podmienenej pravdepodobnosti* výsledku pri danej hypotéze. Nazvime ju *distribúciou podmienenej pravdepodobnosti* pre túto hypotézu, našou šancou vidieť

každý konkrétny výsledok, *ak* je táto hypotéza pravdivá. Distribúcia podmienenej pravdepodobnosti pre hypotézu H je funkcia zložená zo všetkých podmienených pravdepodobností pre p(0 | H) = 0,001, p(1 | H) = 0,001... p(51 | H) = 0,9... p(99 | H) = 0,001. Masa pravdepodobnosti obsahnutá v tejto distribúcii podmienenej pravdepodobnost musí dávať súčet 1. Je to všeobecné pravidlo, že nemôžeme mať šancu 90 %, že uvidíme 51, a zároveň šancu 90 %, že uvidíme 92. Preto, ak najprv predpokladáme, že hypotéza H je pravdivá, stále nie je možné, aby sme mali šancu 90 % vidieť 51, a šancu 90 % vidieť 52.

Táto presná teória predpovedá pravdepodobnosť 90 %, že uvidíme 51. Majme okrem toho aj nejakú teóriu, ktorá predpovedá „pravdepodobnosť 90%, že uvidíme päťdesiat-niečo“.

Ak vidíme výsledok 51, nepovieme, že tento výsledok potvrdzuje obe teórie rovnako. Obe teórie urobili predpoveď, a obe priradili pravdepodobnosť 90 %, a výsledok 51 potvrdzuje obe predpovede. Ale tá presná teória má výhody, pretože sústredila svoju masu pravdepodobnosti do ostrejšieho bodu. Ak tá nejasná teória nič viac neupresnila, počítame „pravdepodobnosť 90 %, že uvidíme päťdesiat-niečo“ ako pravdepodobnosť 9 %, že uvidíme každé konkrétne číslo od 50 do 59.

Predpokladajme, že sme začali s vyrovnanými šancami v prospech presnej i nepresnej teórie – šance 1 : 1, alebo 50 % pravdepodobnosť, že každá z týchto hypotéz je pravdivá. Keď vidíme výsledok 51, aká je výsledná pravdepodobnosť, že je pravdivá tá presná teória? Predpovede týchto dvoch teórií sú analogické ich priradeniu podmienenej pravdepodobnosti – podmienenej pravdepodobnosti, že uvidíme tento výsledok za predpokladu, že táto teória je pravdivá. Aký je pomer podmienených pravdepodobností medzi týmito dvoma teóriami? Prvá teória priradila 90 % masy pravdepodobnosti tomuto *presnému* výsledku. Nejasná teória priradila 9 % masy pravdepodobnosti tomu istému výsledku. Pomer podmienených pravdepodobností je 10 : 1. Ak sme teda začali so šancou 1 : 1, výsledné šance sú 10 : 1 v prospech presnej teórie. Rozdielny tlak týchto dvoch podmienených pravdepodobností posunul našu pôvodnú dôveru 50 % na výslednú dôveru zhruba 91 %, že tá presná teória je správna. *Za predpokladu*, že testujeme iba tieto dve hypotézy, že toto sú jediné indície, ktoré berieme do úvahy, a tak ďalej.

Prečo nejasná teória prehrala, keď obe teórie zodpovedajú indícii? Nejasná teória je plachá; robí širokú predpoveď, chráni si svoje stávky, umožňuje veľa možností, ktoré by falzifikovali tú presnú teóriu. Toto nie je cnosť vedeckej teórie. Filozofi vedy nám hovoria, že teórie by mali byť odvážne, a ochotne sa podrobovať falzifikácii, ak ich predpovede zlyhajú.[221] Teraz vidíme, prečo. Presná teória sústreďuje svoju masu pravdepodobnosti do ostrejšieho bodu a preto sa necháva zraniteľnou falzifikácii, ak skutočný výsledok dopadne niekam inam; ale ak je predpovedaný výsledok správny, presnosť má ohromnú výhodu podmienenej pravdepodobnosti nad nejasnosťou.

Zákony teórie pravdepodobnosti nenechávajú žiadnu možnosť podvádzať, urobiť nejasnú hypotézu tak, aby sa každý výsledok medzi 50 a 59 počítal za rovnako priaznivé potvrdenie ako dostane presná teória, pretože to by vyžadovalo, aby masa pravdepodobnosti dávala dokopy 900 %. Neexistuje spôsob, ako podvádzať, pokiaľ svoje predpovede zaznamenáte *vopred*, takže nemôžete dodatočne tvrdiť, že vaša teória priradila pravdepodobnosť 90 % tomu konkrétnemu výsledku, ktorý nastal. Ľudia majú veľkú záľubu v robení predpovedí dodatočne, takže spoločenský proces vedy vyžaduje predpoveď vopred, skôr než povieme, že nejaký výsledok potvrdzuje nejakú teóriu. Ale ako sa ľudia pohybujú v harmónii s Bayesovou cestou a tak držia jej moc, je téma oddelená od toho, či táto matematika funguje. Keď robíme matematiku, berieme skrátka ako dané, že funkcie hustoty podmienenej pravdepodobnosti sú pevne dané vlastnosti hypotéz, a že masa pravdepodobnosti dáva súčet 1, a nikdy nesnívate o tom, že by ste to robili nejako inak.

Môžete si vyhradiť chvíľku na to, aby ste si názorne predstavili, že *ak* definujeme pravdepodobnosť pomocou kalibrácie, Bayesova veta sa týka kalibrácie. Predstavte si, že si myslím, že Teória 1 má pravdepodobnosť 50 %, že bude pravdivá, a že si myslím, že Teória 2 má pravdepodobnosť 50 %, že bude pravdivá. Predstavte si, že som dobre kalibrovaný; že keď vyslovím slová „päťdesiat percent“, daná udalosť nastane približne v polovici prípadov. A potom vidím výsledok V, ktorý by podľa Teórie 1 nastal asi v deviatich prípadov z desiatich, a podľa Teórie 2 as v deviatich prípadoch zo sto, a použijem

221 Karl R Popper, *The Logic of Scientific Discovery* (New York: Basic Books, 1959).

bayesovské uvažovanie. Ak som bol na začiatku dokonale kalibrovaný (napriek slabému rozlíšeniu, keď som povedal 50/50), stále budem dokonale kalibrovaný (a budem lepšie rozlišovať), keď poviem, že moja dôvery v Teóriu 1 je teraz 91 %. Keby som zopakoval tento druh situácie mnohokrát, mal by som pravdu asi v desiatich prípadoch z jedenástich pri povedaní „91 %“. Ak uvažujem pomocou bayesovských pravidiel, a začnem z dobre kalibrovaného východiska, potom aj moje závery budú dobre kalibrované. Toto platí iba ak definujeme pravdepodobnosť pomocou kalibrácie. Ak namiesto toho slová „istota 90 %“ interpretujeme ako, povedzme, silu našej emócie istoty, potom nie je dôvod očakávať, že výsledná emócia bude v presnom bayesovskom vzťahu v počiatočnou emóciou.

Nech sú pôvodné šance desať k jednej v prospech nejasnej teórie. Prečo? Predpokladajme, že nám náš spôsob opisovania hypotéz dovoľuje zadať buď presné číslo, alebo iba prvú číslicu; môžeme povedať „51“, „63“, „72“, alebo „päťdesiat-niečo“, „šesťdesiat-niečo“, „sedemdesiat-niečo“. Predpokladajme, že si myslíme, že správna odpoveď má asi rovnakú šancu byť odpoveďou toho prvého alebo toho druhého typu. Avšak, keď je daný problém, existuje sto hypotéz toho prvého druhu, ale iba desať hypotéz toho druhého druhu. Ak si teda myslíme, že každá *trieda* hypotéz má asi rovnakú pôvodnú šancu byť správna, potom musíme rozdeliť rovnakú masu pôvodnej pravdepodobnosti medzi desaťkrát toľko presných ako nepresných teórií. Presná teória, ktorá predpovedá presne 51, bude mať teda desatinu pôvodnej masy pravdepodobnosti ako nejasná teória, ktorá predpovedá päťdesiat-niečo. Keď vidíme 51, šance by sa zmenilo z 1 : 10 v prospech nejakej teórie na 1 : 1, vyrovnané šance pre presnú aj nepresnú teóriu.

Ak sa na to pozriete opatrne, je to presne to, čo by zdravý rozum očakával. Začnete s neistotou o tom, či je to ten typ javu, ktorý stále vytvára ten istý výsledok, alebo či je to ten typ javu, ktorý stále vytvára výsledky v nejakom rozsahu. (Možno je tento jav cenový rozsah v supermarkete, ak potrebujete nejaký dôvod predpokladať, že 50..59 je prijateľný rozsah, ale 49..58 nie je.) Urobíte jedno meranie a odpoveď je 51. Nuž, môže to byť preto, že tento jav je presne 51, alebo preto, že je to pätdesiat-niečo. Zostávajúca presná teória má teda rovnakú šancu ako zostávajúca nejasná teória, čo si vyžaduje, aby táto nejasná teória začínala ako desaťkrát pravdepodobnejšia než tá presná teória, keďže presná teória ostrejšie zodpovedá indícii.

Ak vidíme iba jedno číslo, napríklad 51, to nemení pôvodnú pravdepodobnosť, že samotný tento jav je „presný“ alebo „nejasný“. Sústreďuje však celú masu pravdepodobnosti týchto dvoch *tried* hypotéz do jednej prežívajúcej hypotézy z každej triedy.

Samozrejme, je hrubá chyba povedať, že nejaký *jav* je presný alebo nejasný, je to prípad toho, čo Jaynes nazýva Klam projekcie mysle.[222] Presnosť alebo nejasnosť je vlastnosť mapy, nie územia. Namiesto toho by sme sa mali opýtať, či cena v supermarkete zostáva rovnaká, alebo sa trochu mení. Hypotéza „nejasného“ typu je dobrým popisom ceny, ktorá sa trochu mení. Presná mapa by sedela na nemenné územie.

Iný príklad: Hodíte desaťkrát mincou a vidíte postupnosť HHTTH:TTTTH. Možno ste si na počiatku mysleli, že je šanca 1 %, že táto minca je manipulovaná. Priraďuje hypotéza „táto minca je manipulovaná, aby dala HHTTH:TTTTH“ tisíckrát viac masy podmienenej pravdepodobnosti pozorovanému výsledku v porovnaní s hypotézou vyváženej mince? Áno. Posunuli sa výsledné šance, že minca je manipulovaná na 10 : 1? Nie. Pôvodná pravdepodobnosť 1 %, že „minca je manipulovaná“, musí pokrývať všetky možné prípady manipulovanej mince – mincu manipulovanú, aby dala HHTTH:TTTTH, mincu manipulovanú, aby dala TTHHT:HHHHT, atď. Pôvodná pravdepodobnosť, že minca je manipulovaná, aby dala HHTTH:TTTTH, nie je 1 %, ale tisícina z jedného percenta. Až potom, výsledná pravdepodobnosť, že minca je manipulovaná, aby dávala HHTTH:TTTTH, je jedno percento. Čím hovorím: Mysleli ste si, že minca je asi vyvážená, ale že má šancu jedno percento, že je zmanipulovaná, aby dávala nejakú pevne danú náhodne vyzerajúcu postupnosť; hodili ste mincou; minca dala náhodne vyzerajúcu postupnosť; a to vám nepovie nič o tom, či je táto minca vyvážená alebo manipulovaná. Povie vám to, že ak je minca manipulovaná, na *ktorú* postupnosť je manipulovaná.

222 Jaynes, *Probability Theory*.

Toto podobenstvo pomáha vykresliť, prečo Bayesovci *musia* myslieť na pôvodné pravdepodobnosti. Existuje vetva štatistiky, niekedy nazývaná „ortodoxná“ alebo „klasická“ štatistika, ktorá trvá na tom, že pozornosť venujeme iba podmieneným pravdepodobnostiam. Lenže ak venujete pozornosť iba podmieneným pravdepodobnostiam, potom nakoniec nejaká hypotéza o manipulovanej minci vždy porazí hypotézu o vyváženej minci, čo je jav známy ako „overfitting“ teórie na údaje. Po tridsiatich hodoch je *podmienená pravdepodobnost* miliardukrát väčšia pre hypotézu manipulovanej mince než pre hypotézu vyváženej mince. Iba ak je hypotéza manipulovanej mice (alebo skôr, takto konkrétne manipulovanej mince) miliardukrát menej pravdepodobná a priori, môže hypotéza manipulovanej mince prípadne prehrať voči hypotéze vyváženej mince.

Ak potrasiete mincou, aby ste ju resetovali, a začnete hádzať *znova*, a minca dá *znova* HHTTH:TTTTH, to je iná vec. Toto dvíha výslednú šancu manipulovanej mince na 10 : 1, aj keď štartová pravdepodobnosť bola iba 1 %.

Podobne, ak urobíme po sebe dve merania počítadla častíc (alebo cien v supermarkete v stredu) a *obe* merania vrátia 51, presná teória vyhráva šancou 10 k 1.

Presná teória teda vyhráva, ale nejasná teória by stále mala mať lepšie skóre než vôbec žiadna teória. Vezmite si tretiu teóriu, hypotézu nulového poznania alebo *distribúciu maximálnej entropie*, ktorý robí každý výsledok od 0 do 99 rovnako pravdepodobným. Predpokladajme, že uvidíme výsledok 51. Nejasná teória dala lepšiu predpoveď než distribúcia maximálnej entropie – priradila väčšiu podmienenú pravdepodobnosť výsledku, ktorý sme pozorovali. Táto nejasná teória je, doslova, lepšia ako nič. Predpokladajme, že sme začali so šancou 1 : 20 v prospech hypotézy úplnej nevedomosti. (Prečo šanca 1 : 20? Existuje iba jedna hypotéza úplnej nevedomosti, a navyše je to mimoriadne jednoduchá a intuitívna hypotéza. Occamova britva.) Po videní výsledku 51, ktorý nejasná teória predpovedala s pravdepodobnosťou 9 %, oproti 1 % pri úplnej nevedomosti, sa výsledné šance posunú na 10 : 20 alebo 1 : 2. Ak potom uvidíme ďalší výsledok 51, výsledné šance idú na 10 : 2 alebo pravdepodobnosť 83 % pre nejasnú teóriu, za predpokladu, že o žiadnej presnejšej teórii neuvažujeme.

Napriek tomu hanblivosť nejasnej teórie – jej neochota dať *presnú* predpoveď a prijať falzifikáciu pri každom inom výsledku – ju robí zraniteľnou voči odvážnej, presnej teórii. (Samozrejme za predpokladu, že táto presná teória správne uhádne výsledok!) Predpokladajme, že pôvodné šance boli 1 : 10 : 200 pre presnú, nepresnú a nevedomú teóriu – pôvodné pravdepodobnosti 0,5 %, 4,7 % a 94,8 % pre presnú, nepresnú a nevedomú teóriu. Tieto čísla odrážajú našu pôvodnú distribúciu pravdepodobnosti pre *triedy* hypotéz, kde je masa pravdepodobnosti rozdelená pre celé triedy takto: 50 % že sa tento jav hýbe cez všetky čísla, 25 % že sa tento jav posúva v rámci nejakej desiatky, a 25 % že tento jav stále opakuje to isté číslo. 1 hypotéza dokonalej nevedomosti, 10 možných hypotéz desiatkového rozsahu, 100 možných hypotéz opakovania čísla. Preto sú pôvodné šance 1 : 10 : 200 pre presnú hypotézu 51, nejasnú hypotézu „päťdesiat-niečo“, a hypotézu dokonalej nevedomosti.

Po videní výsledku 51, ktorý mal priradené pravdepodobnosti 90 %, 9 % a 1 %, výsledné šance idú na 90 : 90 : 200 = 9 : 9 : 20. Po videní ďalšieho výsledku 51 idú výsledné šance na 810 : 81 : 20, alebo 89 %, 9% a 2 %. Presná teória má teraz prednosť pred nejasnou teóriou, ktorá má zase prednosť pred teóriou nevedomosti.

Teraz si vezmite hlúpu teóriu, ktorá predpovedá pravdepodobnosť 90 %, že uvidíme výsledok od 0 do 9. Hlúpa teória priraďuje skutočnému výsledku 51 pravdepodobnosť 0,1 %. Ak boli šance na začiatku 1 : 10 : 200 : 10 pre presnú, nepresnú, nevedomú a hlúpu teóriu, výsledné šance po videní 51 prvýkrát by boli 90 : 90 : 200 : 1. Hlúpa teória bola falzifikovaná (výsledná pravdepodobnosť 0,2 %).

Je možné mať model taký zlý, že je horší ako nič, ak tento model sústreďuje svoju masu pravdepodobnosti preč od skutočného výsledku, dáva sebaisté predpovede zlých odpovedí. Takáto hypotéza je taká biedna, že prehráva proti hypotéze úplnej nevedomosti. Nevedomosť je lepšia než anti-vedomosť.

Poznámka: V oblasti umelej inteligencie je niekdajšia móda ospevovať úžasnosť náhodnosti. Z času na čas výskumník UI zistí, že keď do niektorého zo svojich algoritmov

pridá šum, tento algoritmus funguje lepšie. Tento výsledok je ohlásený s veľkým nadšením, ktoré sprevádza úlisná chvála tvorivých síl chaosu, nepredvídateľnosti, spontánnosti, nevedomosti čo robí vaša UI, a tak ďalej. (Pozrite si ako príklad The Imagination Engines; podľa svojej vlastnej obchodnej literatúry predávajú zranené a zomierajúce neurónové siete.[223]) Ale aký smutný je algoritmus, ak dokážete *zvýšiť* jeho výkon tým, že mu do medzivýsledkov spracovania vstreknete entropiu? Tento algoritmus musí byť natoľko pomätený, že časť jeho práce ide do sústreďovania masy pravdepodobnosti *preč* od dobrých riešení. Ak injekcia náhody spôsobí spoľahlivé zlepšenie, potom niektorá stránka algoritmu musí byť spoľahlivo horšia než náhoda. Iba v UI dokážu ľudia vymýšľať algoritmy *doslova blbšie než vrece zemiakov*, nakopnúť výsledky trochu smerom k nevedomosti, a potom argumentovať, že šum má liečivú moc.

Predpokladajme, že v našom experimente vidíme výsledky 52, 51, 58. Presná teória dáva tejto spojenej udalosti pravdepodobnosť tisíc k jednej krát 90 % krát tisíc k jednej, kým tá nejasnejšia teória dáva tejto spojenej udalosti pravdepodobnosť 9 % na tretiu, čo je dokopy... hm... hm... pozrime sa na to... milión k jednej pre presnú teóriu, verzus tisíc k jednej pre nejasnú teóriu. Alebo tak nejako; počítame tu približné mocniny desiatky. Verzus milión k jednej pre distribúciu nulovej vedomosti, ktorá prideľuje všetkým výsledkom rovnakú pravdepodobnosť. Verzus miliarda k jednej pre model horší ako nič, pre hlúpu hypotézu, ktorá tvrdí, že s pravdepodobnosťou 90 % uvidíme číslo menšie ako 10. Podľa týchto približných čísel získa nejasná teória skóre -30 decibelov (pravdepodobnosť 1/1000 celkového výsledku experimentu), verzus -60 pre presnú teóriu, -60 pre nevedomú teóriu, a -90 pre hlúpu teóriu. Nie je vždy pravda, že vyhrá to najvyššie skóre, pretože musíme vziať do úvahy naše pôvodné šance 1 : 10 : 200 : 10, čo je dôvera -23, -13, 0 a -13 decibelov. Nejasná teória stále vychádza s najvyšším skóre -43 decibelov. (Keby sme ignorovali svoje pôvodné pravdepodobnosti, potom by každý nový experiment mohol prevalcovať akumulované výsledky všetkých predchádzajúcich experimentov; nemohli by sme akumulovať poznanie. Navyše, vždy by vyhrala hypotéza manipulovanej mince.)

Ako vždy, nemali by ste sa znepokojovať, že ešte aj tá najlepšia teória má nízke skóre – spomeňte si na podobenstvo s vyváženou mincou. Teórie sú aproximácie. V princípe by sme mohli byť schopní predpovedať presnú postupnosť hodov mincou. Ale chcelo by to lepšie meranie a viac výpočtovej sily než sme ochotní vynaložiť. Možno by sme s dosť dobrým modelom mohli dosiahnuť predpovede hodu mincou 60/40...? Používame najlepšiu aproximáciu, ktorú máme, a pokúšame sa dosiahnuť dobrú kalibráciu aj keď rozlišovanie nie je dokonalé.

* * *

Našu analýzu sme zatiaľ robili podľa pravidiel bayesovskej teórie pravdepodobnosti, kde neexistuje spôsob, ako mať viac než 100 % masy pravdepodobnosti, a teda žiaden spôsob, ako podvádzať, aby sa hocijaký výsledok mohol počítať ako „potvrdenie“ vašej teórie. Podľa Bayesovho zákona, hracie peniaze nemožno sfalšovať; máte iba obmedzené množstvo hliny.

Nanešťastie, ľudia nie sú Bayesiáni. Ľudia sa čudesne pokúšajú *obhajovať* hypotézy, robia vedomé úsilie dokázať ich alebo zabrániť vyvráteniu. Toto správanie nemá žiadnu analógiu v zákonoch teórie pravdepodobnosti ani teórie rozhodovania. Vo formálnej teórii pravdepodobnosti hypotéza *existuje*, indícia *existuje*, a buď je hypotéza potvrdená alebo nie. Vo formálnej teórii rozhodovania môže činiteľ vynaložiť snahu preskúmať nejakú tému, ktorou si momentálne nie je istý, nevie, či indície pôjdu jedným alebo druhým smerom. V žiadnom prípade sa vedome nepokúša nejakú myšlienku dokázať alebo vyvrátiť. Môže *testovať* myšlienky, ktorými si naozaj nie je istý, ale nemôže mať „uprednostňovaný“ výsledok skúmania. Nemôže sa snažiť hypotézu dokázať, ani zabrániť jej dôkazu. Nedokážem primerane sprostredkovať, aká bláznivá by takáto predstava bola pre skutočného Bayesiána; v bayesovskom jazyku ani neexistujú slová, ktoré by túto chybu opísali...

223 Imagination Engines, Inc., „The Imagination Engine® or Imagitron™,“ 2011, http://www.imagination-engines.com/ie.htm.

Pre každé očakávanie indície existuje rovnaké očakávanie protiindície v opačnom smere. Ak je A indíciou v prospech B, potom nie-A *musí* byť indíciou v prospech nie-B. Sily týchto indícií nemusia byť rovnaké; zriedkavá ale silná indícia jedným smerom môže byť vyvážená častou alebo slabou indíciou opačným smerom. Ale nie je možné, aby aj A aj nie-A boli indíciou v prospech B. To jest, nie je to možné podľa zákonov teórie pravdepodobnosti.

Ľudia často chcú nechať si svoj koláč a zároveň ho zjesť. Hocijaký výsledok, ktorý vidíme, je ten, ktorý potvrdzuje našu teóriu. Ako povedal Spee, kňaz z kapitoly Zákon zachovania očakávanej indície: „Vyšetrujúca komisia by sa cítila zneuctená, keby nejakú ženu oslobodila; keď raz bola uväznená a v reťaziach, musela byť uznaná vinnou, či už poctivo alebo nepoctivo.“[224]

Zdá sa, že ľudská psychológia funguje tak, že najprv vidíme, ako sa niečo stane, a potom sa snažíme argumentovať, že to zodpovedá ľubovoľnej hypotéze, ktorú sme predtým mali v hlave. Namiesto zachovávania masy pravdepodobnosti, ktorú by sme vopred rozdeľovali medzi *predpovede*, máme pocit *súladu* – stupeň, nakoľko vysvetlenie a udalosť napohľad do seba „zapadajú“. Toto „zapadanie“ sa nezachováva. Neexistuje žiaden ekvivalent pravidla, že masa pravdepodobnosti musí dávať súčet 1. Psychoanalytik môže vysvetliť ľubovoľné možné správanie pacienta zostavením príslušnej štruktúry „racionalizácií“ a „obrán“; zapadá to, preto to musí byť pravda.

Teraz si predstavte príbeh z kapitoly Falošné vysvetlenia – študenti vidia radiátor a vedľa radiátora kovový plát. Študenti by nikdy nepredpovedali vopred, že strana plátu pri radiátore bude chladnejšia. Napriek tomu, keď videli tento fakt, podarilo sa im dosiahnuť, aby ich vysvetlenia „zapadli“. Stratili svoju vzácnu šancu na zmätok, na uvedomenie si, že ich modely nepredpovedajú jav, ktorý pozorovali. Obetovali svoju schopnosť byť viac zmätení fikciou než pravdou. A neuvedomili si, že „tepelná indukcia, bla bla, preto je blízka strana chladnejšia“ je nejasná a slovná predpoveď, roztiahnutá cez nenormálne široký rozsah možných hodnôt konkrétnych nameraných teplôt. Použitie rovníc rozptylu a rovnováhy by dalo *ostrú* predpoveď možných spoločných hodnôt. Možno by neupresnilo tie *prvé* hodnoty, ktoré by ste namerali, ale keby ste vedeli pár hodnôt, mohli by ste vytvoriť ostrú predpoveď pre tie zvyšné. Skóre pre celkový experimentálny výsledok by bolo omnoho lepšie než ľubovoľná menej presná alternatíva, najmä nejasná a slovná predpoveď.

* * *

Teraz máte *technické* vysvetlenie rozdielu medzi slovným vysvetlením a technických vysvetlením. Je to technické vysvetlenie, pretože vám umožňuje spočítať, *nakoľko presne* je nejaké vysvetlenie technické. Nejasné hypotézy môžu byť také nejasné, že iba nadľudská inteligencia by dokázala spočítať, ako presne nejasné sú. Dostatočne obrovská inteligencia by azda dokázala extrapolovať všetky možné experimentálne výsledky, extrapolovať všetky možné závery nejasne hádajúceho o tom, ako dobre táto nejasná hypotéza „zapadá“, a potom renormalizovať túto distribúciu „zapadania“ na distribúciu podmienenej pravdepodobnosti, ktorej súčet by bol 1. Ale v princípe je stále možné presne spočítať, nakoľko nejasná je nejasná hypotéza. Ten výpočet akurát nie je výpočtovo zvládnuteľný, rovnako ako počítanie dráhy lietadla pomocou kvantovej mechaniky nie je výpočtovo zvládnuteľné.

Trvám na tom, že každý by sa mal naučiť aspoň jeden technický predmet. Fyziku; informatiku; evolučnú biológiu; bayesovskú teóriu pravdepodobnosti, ale aspoň *niečo*. Niekto, kto nemá v rukáve *žiaden* technický predmet, nemá žiaden referent pre to, čo to znamená niečo „vysvetliť“. Môže si myslieť, že „Všetko je oheň“ je vysvetlenie, ako si to myslel grécky filozof Hérakleitos. Preto zastávam postoj, že by sa bayesovská teória pravdepodobnosti mala vyučovať na strednej škole. Bayesovská teória pravdepodobnosti je jediný kus matematiky, o ktorom viem, že je dostupný na stredoškolskej úrovni, a ktorý umožňuje *technické* pochopenie témy – dynamiky názoru – ktorá patrí do oblasti každodeného života a má emocionálne významné dôsledky. Študovať bayesovskú pravdepodobnosť by dalo študentom referent pre to, čo znamená niečo „vysvetliť“.

Príliš veľa akademikov si myslí, že byť „technický“ znamená hovoriť suchými mnohoslabičnými slovami. Tu je „technické“ vysvetlenie technického vysvetlenia:

224 Friedrich Spee, *Cautio Criminalis; or, A Book on Witch Trials*, ed. and trans. Marcus Hellyer, Studies in Early Modern German History (1631; Charlottesville: University of Virginia Press, 2003).

> Rovnice teórie pravdepodobnosti uprednostňujú hypotézy, ktoré silno predpovedajú presné pozorované údaje. Silné modely odvážne sústredia svoju hustotu pravdepodobnosti do presných výsledkov, čo ich robí falzifikovateľnými, ak dáta dopadnú inde, a dáva im ohromnú výhodu v podmienenej pravdepodobnosti oproti menej odvážnym, menej presným modelom. Slovné vysvetlenia fungujú na psychologickom vyhodnocovaní nezachovávaného zapadania post facto namiesto zachovávanej hustoty pravdepodobnosti ante facto. A slovné vysvetlenie nevykresľuje ostro podrobné obrázky, čím naznačuje plynulú distribúciu pravdepodobnosti v okolí údajov.

Je toto uspokojujúce? Nie. Vypočute si tieto pôsobivé a vážne vety, znejúce s tupým buchotom odbornosti. Pozrite si nešťastných študentov, ako tieto vety píšu na hárky papiera. Ešte aj potom, čo si poslucháči vypočuli tieto rituálne slová, nedokážu vykonať žiaden výpočet. *Vy* tú matematiku poznáte, takže vám tie slová dávajú zmysel. Vy dokážete vykonať výpočty po tom, čo ste si vypočuli tieto pôsobivé slová, rovnako ako ste to mohli urobiť aj predtým. Ale čo ten, kto nevidel urobiť žiadne výpočty? Aké nové zručnosti získal z tejto „technickej“ prednášky, okrem schopnosti recitovať fascinujúce slová?

„Bayesovský“ je iste fascinujúce slovo, nie? Poďme ho dostať von zo svojho systému: Bayes Bayes Bayes Bayes Bayes Bayes Bayes Bayes Bayes...

Posvätná slabika je nezmyselná, okrem toho, ak niekomu hovorí, aby použil matematiku. Preto ten, kto počúva, musí najprv poznať matematiku.

A naopak, ak už viete matematiku, môžete byť takí pochabý, ako sa vám páči, a stále ste technický.

Takto sa teda zbavujeme opäť ďalšieho stereotypu rozumnosti, že rozumnosť sa skladá zo suchej formálnosti a vážnosti bez humoru. Čo má toto spoločné s problémom rozlíšenia pravdy od nepravdy? Čo má toto spoločné s hľadaním mapy, ktorá odráža územie? Vedec hodný laboratórneho plášťa by mal byť schopný robiť originálne objavy aj keby nosil šašovský oblek, alebo prednášať vysokým piskľavým hlasom po vdychovaní hélia. Nie je napísané nikde v matematike teórie pravdepodobnosti, že človek sa nesmie zabávať. Čepeľ, ktorá tne až k správnej odpovedi, nemá v sebe žiadnu dôstojnosť ani pochabosť, hoci môže sedieť v ruke pochabého vládcu.

* * *

Užitočný model nie je jednoducho niečo, čo viete, tak ako viete, že lietadlo sa skladá z atómov. Užitočný model je poznanie, ktoré môžete spočítať v rozumnom čase, aby ste predpovedali udalosti skutočného sveta, ktoré viete ako pozorovať. Možno niekto zistí, že použitím modelu, ktorý trošku porušuje zákon zachovania hybnosti, dokážete spočítať aerodynamiku lietadla 747 omnoho *lacnejšie* než keby ste trvali na tom, aby sa hybnosť presne zachovávala. Ak by ste teda nechali dva počítače súťažiť o najlepšiu predpoveď, mohlo by to byť tak, že tá najlepšia predpoveď príde z modelu, ktorý porušuje zachovanie hybnosti. To neznamená, že lietadlo 747 porušuje zachovanie hybnosti v skutočnom živote. Žiaden model nepoužíva jednotlivé atómy, ale z toho nevyplýva, že lietadlo 747 sa neskladá z atómov. Fyzici používajú rôzne modely na predpovedanie lietadiel a zrážok častíc, pretože by bolo príliš drahé počítať lietadlo časticu po častici.

Mohli by ste dokázať, že lietadlo 747 sa skladá z atómov pomocou experimentálnych údajov, ktoré aerodynamické modely nespracovávajú; napríklad by ste mohli nacvičiť skenujúci tunelový mikroskop na časti krídla a pozrieť sa na tie atómy. Podobne by ste mohli použiť jemnejší merací nástroj na rozlíšenie medzi lietadlom 747, ktoré *naozaj* porušuje zachovanie hybnosti ako predpovedá táto lacná aproximácia, a lietadlom 747, ktoré dodržiava zachovanie hybnosti, ako predpovedá základná fyzika. Víťazná teória je tá, ktorá najlepšie predpovedá všetky experimentálne predpovede dokopy. Naše bayesovské bodovacie pravidlo nám dáva spôsob, ako skombinovať výsledky *všetkých* našich experimentov, dokonca aj experimentov, ktoré používajú rôzne metódy.

Navyše, atómová teória povoľuje, zahŕňa a v istom zmysle splnomocňuje aerodynamický model. Ak myslíme abstraktne na predpoklady atómovej teórie, uvedomíme si, že aerodynamický model by mal byť dobrou (a omnoho lacnejšou) aproximáciou atómovej teórie, a tak atómová teória podporuje

aerodynamický model, namiesto aby mu konkurovala. Úspešná teória môže zahŕňať mnoho modelov pre rôzne oblasti, dokiaľ tieto modely uznávame ako aproximácie, a dokiaľ je v každom prípade daný model kompatibilný s (ideálne požadovaný) základnou teóriou.

Naša *základná* fyzika – kvantová mechanika, štandardné častice, a relativita – je teória, ktorá zahŕňa *ohromnú* rodinu modelov pre makroskopické fyzikálne javy. Existuje fyzika tekutín a pevných telies a plynov; toto však neznamená, že vo svete existujú *základné* veci, ktoré majú vnútornú vlastnosť tekutosti.

> Zdanlivo existuje farba, zdanlivo sladká chuť, zdanlivo horká chuť, v skutočnosti sú iba atómy a prázdnota.[225]
>
> --Demokritus, 420 p.n.l. (podľa Robinson a Groves)

* * *

Argumentovaním, že „technická“ teória by mala byť definovaná ako teória, ktorá ostro sústreďuje pravdepodobnosť do konkrétnych predpovedí vopred, nastavujem extrémne vysoká latku prísnosti. Videli sme, že nejasná teória *môže* byť lepšia ako nič. Nejasná teória môže vyhrať nad hypotézou nevedomosti, ak neexistujú presné teórie, ktoré by jej konkurovali.

Existuje ohromná rodina modelov patriacich do centrálnej základnej teórie života a biológie; táto základná teória sa niekedy volá neodarwinizmus, prírodzený výber alebo evolúcia. Niektoré modely v evolučnej teórii sú kvantitatívne. Spôsob, akým DNA kóduje bielkoviny je nadbytočný; dve rôzne postupnosti DNA môžu kódovať presne tú istú bielkovinu. Existujú 4 bázy DNA {ATCG} a 64 možných kombinácií troch báz DNA. Ale týchto 64 možných kodónov opisuje iba 20 aminokyselín plus značku stop. Genetický drift by teda mal produkovať nefunkcionálne zmeny v genómoch druhov, pomocou mutácií, ktoré sa náhodou zafixujú v genofonde. Miera akumulácie nefunkcionálnych rozdielov medzi genómami dvoch druhov so spoločným predkom závisí od takých parametrov, ako je počet uplynulých generácií a intenzita výberu v tomto genetickom mieste. Toto je príklad člena rodiny evolučných modelov, ktorý produkuje kvantitatívne predpovede. Existuje aj nerovnováha frekvencie alel pri výbere, stabilné rovnováhy stratégií teórie hier, pomer pohlaví, a tak ďalej.

Toto všetko patrí pod hlavičku „fascinujúce slová“. Nanešťastie existujú isté náboženské frakcie, ktoré šíria hrubé dezinformácie o evolučnej teórii. Preto zdôrazňujem, že mnohé modely v evolučnej teórii robia kvantitatívne predpovede, ktoré sú experimentálne potvrdené, a že takéto modely viac než stačia na ukážku, že napríklad ľudia a šimpanzy majú spoločného predka. Ak ste obeťou dezinformácie kreacionistov – to jest, ak ste počuli nejaké náznaky, že evolučná teória je kontroverzná alebo netestovateľná alebo „iba teória“ alebo nie dôkladná alebo nie technická alebo nejako inak nie potvrdená nepredstaviteľne obrovskou hromadou experimentálnych indícií – odporúčam vám prečítať si *TalkOrigins FAQ*[226] a študovať evolučnú biológiu s matematikou.

Ale predstavte si, že sa vrátite späť v čase do devätnásteho storočia, keď teóriu prirodzeného výberu práve objavil Charles Darwin a Alfred Russel Wallace. Predstavte si evolucionizmus tesne po jeho vzniku, keď táto teória nemala nič, čo by sa podobalo modernému súboru kvantitatívnych modelov a veľkým rastúcim horám experimentálnych indícií. Nedalo sa nijako vedieť, že objavíme, že ľudia a šimopanzy majú spoločných 95 % genetického materiálu. Nikto nevedel, že existuje DNA. Napriek tomu sa vedci zhromaždili okolo novej teórie prirodzeného výberu. A neskôr sa ukázalo, že *existuje* presne kopírovaný genetický materiál s potenciálom mutovať, že ľudia a šimpanzy sú dokázateľne príbuzní, atď.

Takže tento veľmi prísny, veľmi vysoký štandard, ktorý som navrhol pre „technickú“ teóriu je príliš prísny. Historicky *bolo* možné úspešne rozlíšiť pravdivé teórie od nepravdivých teórií na základe predpovedí toho typu, ktorý nazývam „nejasné“. Nejasné predpovede s istotou povedzme 80 % si dokážu vybudovať veľkú výhodu voči alternatívnym hypotézam, ak je dostatok experimentov. Azda by sa teória tohto druhu, produkujúca predpovede, ktoré nie sú presne podrobné, ale napriek tomu sú správne, dala nazvať „polotechnická“?

---

225 Citované v Dave Robinson and Judy Groves, *Philosophy for Beginners*, 1st ed. (Cambridge: Icon Books, 1998).

226 TalkOrigins Foundation, „Frequently Asked Questions about Creationism and Evolution,“ http://www.talkorigins.org/origins/faqs-qa.html.

Ale technické teórie sú iste spoľahlivejšie než polotechnické teórie? Technické teórie by iste mali mať prednosť, zaslúžiť si väčšiu úctu? Fyzika, ktorá produkuje mimoriadne presné predpovede, je iste v nejakom zmysle lepšie potvrdená ako evolučná teória? Čím samozrejme nenaznačujem, že evolučná teória je nesprávna; ale nech sú hory indície v prospech evolúcie akokoľvek rozsiahle, nie je na tom fyzika lepšie s rozsiahlymi horami *presných* experimentálnych potvrdení? Pozorovanie neutrónových hviezd potvrdzujú predpovede všeobecnej relativity s presnosťou jedna ku sto biliónom (10^{14}). Má evolučná teória niečo, čo by zodpovedalo tomuto?

Daniel Dennett raz povedal, že ak meriame podľa jednoduchosti teórie a množstva zložitosti, ktoré vysvetľuje, Darwin mal tú najúžasnejšiu myšlienku v celej histórii.[227]

Kedysi existoval konflikt medzi fyzikou 19. storočia a evolucionizmom 19. storočia. Podľa najlepších vtedajších fyzikálnych modelov, Slnko nemohlo horieť príliš dlho. 3000 rokov pomocou chemickej energie, alebo 40 miliónov rokov podľa gravitačnej energie. Fyzika 19. storočia nepoznalo žiaden zdroj energie, ktorý by umožnil horieť dlhšie. Fyzika 19. storočia nebola *až* taká mocná ako moderná fyzika – nemala predpovede presné na jedna ku 10^{14}. Ale fyzika 19. storočia predsa mala matematickú povahu modernej fyziky; disciplína, ktorej modely dávali podrobné, presné, kvantitatívne predpovede. Evolučná teória 19. storočia bola celkom polotechnická, bez štipy kvantitatívneho modelovania. Dokonca ani Mendelove experimenty s hrachom vtedy neboli známe. A predsa vyzeralo pravdepodobne, že evolúcia by na svoje fungovanie potrebovala viac než biednych 40 miliónov rokov – stovky miliónov, možno miliardy rokov. Starobylosť Zeme bola nejasnou a polotechnickou predpoveďou nejasnej a polotechnickej teórie. Oproti tomu mali fyzici 19. storočia presné a kvantitatívne modely, ktoré pomocou formálnych výpočtov dali presný a kvantitatívny rozsudok, že Slnko jednoducho nemohlo horieť tak dlho.

> Obmedzenia geologických období dané fyzikou nemôžu samozrejme vyvrátiť hypotézu premeny druhov; ale zdajú sa dostatočné na vyvrátenie doktríny, že táto premena nastala vďaka „dedičnosti s úpravami podľa prirodzeného výberu."
>
> --Lord Kelvin, podľa Lyle Zapato[228]

História zaznamenáva, kto vyhral.

Ponaučenie? Ak dokážete dať s istotu 80 % predpovede vopred na otázky typu „áno alebo nie", môže to byť „nejasná teória", môže sa mýliť v jednom prípade z piatich, a stále môžete vybudovať sakra veľký náskok v skóre oproti hypotéze nevedomosti. Dosť na to, aby to teóriu potvrdilo, ak neexistujú lepší konkurenti. Skutočnosť je konzistentná; každá *správna* teória o vesmíre je zlučiteľná a každou inou správnou teóriou. Nedokonalé mapy si môžu odporovať, ale existuje iba jedno územie. Evolucionizmus 19. storočia mohol byť polotechnickým odvetvím, ale bol správnym (ako dnes vieme) a zďaleka najlepším vysvetlením (dokonca aj za oných čias). Ľubovoľný konfikt medzi evolucionizmom a inou dobre potvrdenou teóriou musel odrážať nejaký druh anomálie, chybu v tvrdení, že tieto dve teórie sú nezlučiteľné. Fyzika 19. storočia nemohla modelovať dynamiku Slnka – nepoznala jadrové reakcie. Nemohla ukázať, že jej chápanie Slnka bolo správne *v technických detailoch,* ani počítať na základe *potvrdeného* modelu Slnka, aby určila, ako dlho Slnko existovalo. V spätnom pohľade môžeme teda povedať niečo ako: „Bola tam možnosť, že fyzika 19. storočia jednoducho nechápala Slnko."

Ale to je spätný pohľad. Skutočné ponaučenie je, že aj keď bola fyzika 19. storočia presná a kvantitatívna, nevyhrala automaticky nad polotechnickou teóriou evolucionizmu 19. storočia. *Obe* tieto teórie boli dobre podporené. *Obe* boli správne v oblastiach, nad ktorými zovšeobecňovali. Zdanlivý konflikt medzi nimi bol anomáliou, a ukázalo sa, že táto anomália korení v neúplnosti a nesprávnom aplikovaní fyziky 19. storočia, nie v neúplnosti a nesprávnom použití evolucionizmu 19. storočia. Ale bolo by márne porovnávať horu indícií podporujúcich jednu teóriu s horou indícií podporujúcich druhú. Ešte aj za oných čias boli obe hory príliš veľké na to, aby sme predpokladali, že sa niektoré z teórií

227 Daniel C. Dennett, *Darwin's Dangerous Idea: Evolution and the Meanings of Life* (Simon & Schuster, 1995).

228 Citované v Lyle Zapato, „Lord Kelvin Quotations," 2008, http://zapatopi.net/kelvin/quotes/.

jednoducho mýli. Takéto veľké hory indície nemôžete nechať súťažiť proti sebe, ako keby jedna falzifikovala druhú. Musí to byť tak, že jednu teóriu aplikujete nesprávne, alebo ju aplikujete mimo oblasti, v ktorej predpovedá dobre.

Nemali by ste sa teda *nutne* vysmievať nejakej teórii len preto, že je polotechnická. Polotechnické teórie dokážu nahromadiť dostatočne vysoké skóre v porovnaní s každou dostupnou alternatívou, takže viete, že táto teória je aspoň približne správna. Jedného dňa môže byť táto polotechnická teória nahradená alebo dokonca falzifikovaná nejakou úspešnejšou konkurenciou, ale to platí aj pre technické teórie. Pomyslite na to, ako Einsteinova všeobecná relativita zožrala Newtonovu gravitačnú teóriu.

Ale správnosť polotechnickej teórie – teórie, ktorá momentálne nemá žiadne presné výpočtovo zvládnuteľné modely testovateľné vykonateľnými experimentmi – môže byť omnoho menej jasná než správnosť technickej teórie. Stále treba zručnosť, trpezlivosť a skúmanie na rozlíšenie dobrých polotechnických teórií od teórií, ktoré sú iba celkom zmätené. Toto nie je niečo, čo by ľudia dobre robili pomocou inštinktu, a preto máme Vedu.

Ľudia radi skočia na návnadu a chopia sa ľubovoľného dostupného dôvodu, prečo odmietnuť teóriu, ktorá sa im nepáči. To je dôvod, prečo som dal príklad evolucionizmu 19. storočia, aby som ukázal, prečo by človek nemal príliš rýchlo spakruky odmietnuť „netechnickú“ teóriu. Podľa morálnych zvykov vedy bol evolucionizmus 19. storočia vinný z nejedného hriechu. Evolucionizmus 19. storočia nedáva žiadne kvantitatívne predpovede. Nebol pripravený na falzifikáciu. Bolo to prevažne vysvetlenie toho, čo už sme videli. Chýbal mu základný mechanizmus, lebo vtedy nikto nevedel o DNA. Bol dokonca v rozpore so zákonmi fyziky 19. storočia. Napriek tomu bol prirodzený výber takým *úžasne dobrým* vysvetlením post facto, že sa k nemu ľudia zbehli, a potom sa ukázalo, že je to správne. Veda, ako ľudská aktivita, si vyžaduje predpovede vopred. Teória pravdepodobnosti, ako matematika, nerozlišuje medzi predpoveďami post facto a vopred, pretože teória pravdepodobnosti predpokladá, že distribúcia pravdepodobnosti je pevne daná vlastnosť hypotézy.

Pravidlo o predpovediach vopred je pravidlom spoločenského procesu vedy – je to morálny zvyk, nie veta. Tento morálny zvyk existuje preto, aby zabránil ľuďom robiť ľudské chyby, ktoré by sa v jazyku teórie pravdepodobnosti ťažko dali opísať, ako napríklad dodatočné dolaďovanie, čo vlastne tvrdíte, že vaša hypotéza predpovedá. Ľudia došli k záveru, že evolucionizmus 19. storočia je vynikajúce vysvetlenie, hoci je post facto. Toto uvažovanie bolo *podľa teórie pravdepodobnosti správne*, preto *fungovalo* napriek všetkým vedeckým hriechom. Teória pravdepodobnosti je matematika. Spoločenský proces vedy je sada legálnych zvyklostí, aby sa ľuďom zabránilo v tejto matematike podvádzať.

Je však zároveň pravda, že v porovnaní s *modernými* teoretikmi evolúcie, teoretici evolúcie konca 19. a začiatku 20. storočia často smutne blúdili. Darwin, ktorý bol dosť bystrý na to, aby vynašiel túto teóriu, určil úžasné množstvo vecí správne. Ale Darwinovi nasledovníci, ktorí boli iba dosť bystrí na to, aby túto teóriu prijali, sa v evolúcii mýlili často a vážne. Na opravenie ich chýb bol potom potrebný zvyčajný proces vedy. Je neuveriteľné ako málo chýb v uvažovaní urobil Darwin[229] v knihách *Pôvod druhov* a *Pôvod človeka*, v porovnaní s tými, ktorí ho nasledovali.

Aj toto je nebezpečenstvo polotechnickej teórie. Aj potom, čo sa záblesk geniálneho vhľadu potvrdí, púhi priemerní vedci môžu v neprítomnosti formálnych modelov tieto vhľady aplikovať nesprávne. Ešte koncom 1960-tych rokov biológovia hovorili o tom, ako evolúcia funguje „pre dobro druhu“ alebo naznačovali, že jednotlivci by obmedzili svoje rozmnožovanie, aby zabránili druhu premnožiť sa na nejakom území. Tí najlepší evoluční teoretici vedeli, že to tak nie je, ale priemerní teoretici nie.[230]

229 Charles Darwin, *On the Origin of Species by Means of Natural Selection; or, The Preservation of Favoured Races in the Struggle for Life*, 1st ed. (London: John Murray, 1859), http://darwin-online.org.uk/content/frameset?viewtype=text&itemID=F373&pageseq=1; Charles Darwin, *The Descent of Man, and Selection in Relation to Sex*, 2nd ed. (London: John Murray, 1874), http://darwin-online.org.uk/content/frameset?itemID=F944&viewtype=text&pageseq=1.

230 Williams, *Adaptation and Natural Selection*.

Takže je *omnoho* lepšie mať technickú teóriu než polotechnickú teóriu. Nanešťastie Príroda nie je vždy taká milá, aby sa ukázala v podobe opísateľnej úhľadnými, formálnymi, *výpočtovo zvládnuteľnými* modelmi, ani vždy neposkytuje svojim študentom meracie nástroje, ktorými môžu priamo preskúmať jej javy. Niekedy je to iba otázka času. Evolucionizmus 19. storočia bol polotechnický, ale neskôr prišla matematika populačnej genetiky, a nakoniec sekvenovanie DNA. Ale príroda vám nedáva vždy jav, ktorý môžete opísať pomocou technických modelov pätnásť sekúnd potom, čo máte základný vhľad.

Avšak najnovšia veda, *kontroverzia*, sa najčastejšie týka nejakej polotechnickej teórie, alebo nezmyslu, ktorý sa tvári ako polotechnická teória. V čase, keď teória dosiahne technické postavenie, zvyčajne už nie je kontroverzná (medzi vedcami). Takže otázka, ako odlíšiť dobré polotechnické teórie od nezmyslov je veľmi dôležitá pre vedcov, a nie je to také ľahké, ako zavrhnutie spakruky každej teórie, ktorá nie je technická. Za účelom odlišovania pravdy od nepravdy existuje celá disciplína rozumnosti. Toto umenie sa nedá redukovať na zoznam položiek, alebo aspoň nie na zoznam položiek, ktoré by priemerný vedec dokázal spoľahlivo aplikovať po hodine výcviku. Keby to bolo také jednoduché, nepotrebovali by sme vedu.

* * *

Prečo sa zaujímate o vedecké kontroverzie? Načo sa živiť takou riedkou a hnilou potravou, akú ponúkajú médiá, keď existuje tak veľa kvalitného jedla, ktoré nájdete v učebniciach? Veda učebníc je krásna! Veda učebníc je *zrozumiteľná*, na rozdiel od púhych fascinujúcich slov, ktoré nikdy nemôžu byť naozaj krásne. Fascinujúce slová nemajú žiadnu moc, dokonca ani žiaden zmysel, bez matematiky. Fascinujúce slová nie sú poznanie, ale ilúzia poznania, čo je dôvod, prečo prináša tak málo uspokojenia vedieť, že „gravitácia je výsledkom zakrivenia časopriestoru“. Veda nie je vo fascinujúcich slovách, hoci to je všetko, o čom sa dočítate v čerstvých novinách.

Môže existovať dôvod sledovať vedeckú kontroverziu. Mohli by ste byť odborníkom v danej oblasti, a v tom prípade je táto vedecká kontroverzia vašou správnou stravou. Alebo táto vedecká kontroverzia môže byť niečo, čo potrebujete vedieť *teraz*, pretože to ovplyvňuje váš život. Možno je 19. storočie a vy žiadostivo hľadíte na osobu osobu toho správneho pohlavia oblečenú v plavkách 19. storočia, a potrebujete vedieť, či vaša sexuálna túžba pochádza z psychológie zostavenej prirodzeným výberom, alebo je to pokušenie, ktorému vás vystavil Diabol, aby vás zvábil do pekelného ohňa.

Nie je celkom nemožné natrafiť na vedeckú kontroverziu, ktorá sa nás týka, a zistiť, že pálčivo a súrne potrebujeme správnu odpoveď. Budem preto hovoriť o niektorých varovných znameniach, ktoré historicky rozlišovali nejasné hypotézy, ktoré sa neskôr ukázali ako nevedecký blábol od nejakých hypotéz, ktoré neskôr vyrástli na potvrdené teórie. Spomeňte si len na historickú lekciu evolucionizmu 19. storočia, a odolajte pokušeniu zavrhnúť každú teóriu, ktorej chýba jedna položka na vašom zozname vlastností. Nemám v úmysle dať ľuďom ďalšiu výhovorku na odmietanie dobrej vedy, ktorá im nevyhovuje. Ak na teórie, ktoré sa vám nepáčia, aplikujete prísnejšie kritériá než na teórie, ktoré sa vám páčia (alebo naopak!), potom každý ďalší detail, ktorý sa naučíte vidieť, každá nová logická chyba, ktorú sa naučíte odhaliť, vás urobí o toľko hlúpejším. Inteligencia, aby bola užitočná, musí byť použitá na niečo iné než na porazenie seba samej.

* * *

Jedným z klasických príznakov slabej hypotézy je, že musí vynakladať veľké úsilie, aby sa vyhla falzifikácii – komplikované dôvody, prečo je táto hypotéza zlučiteľná s daným javom, hoci sa daný jav nespráva tak, ako sa očakávalo. Carl Sagan dáva príklad niekoho, kto tvrdí, že v jeho garáži žije drak. Sagan pôvodne vyvodil ponaučenie, že slabé hypotézy potrebujú rýchlo utekať, aby sa vyhli falzifikácii – aby si udržali dojem, že „zapadajú“.[231]

Ja by som podotkol, že dotyčný má zrejme *niekde* vo svojej hlave dobrý model situácie, pretože dokáže predpovedať, v predstihu, presne aké výhovorky bude potrebovať. Pre Bayesovca hypotéza nie je niečo, čo vyslovujete hlasným, pôsobivým tónom. Hypotéza je niečo, čím sa riadia vaše *očakávania*, pravdepodobnosti, ktoré prisudzujete budúcim vnemom. To je to, čo *je* pre Bayesovca pravdepodobnosť

231 Carl Sagan, *The Demon-Haunted World: Science as a Candle in the Dark*, 1st ed. (New York: Random House, 1995).

– to je to, čo bodujete, čo je to, čo kalibrujete. Takže hoci náš dotyčný hovorí hlasno, pôsobivo, a úprimne, že *verí*, že v garáži je neviditeľný drak, *neočakáva*, že v garáži je neviditeľný drak – očakáva presne rovnaké vnemy ako skeptik.

Keď ja hodnotím predpovede hypotézy, pýtam sa, ktoré vnemy by som očakával, nie v ktoré fakty by som veril.

A z opačnej strany:

Nedávno som sa hádal so svojím kamarátom ohľadom evolučnej teórie. Kamarát tvrdil, že zhluky zmien vo fosílnom zázame (zdá sa, že existujú obdobia relatívneho pokoja, po ktorých nasledujú relatívne ostré zmeny; toto samotné je kontroverzné pozorovanie známe ako „prerušovaná rovnováha“) ukazujú, že v našom chápaní vzniku druhov je niečo nesprávne. Môj kamarát si myslel, že tam pôsobí nejaká neznáma sila - nie nadprirodzená, ale niečo prirodzené, čo štandardná evolučná teória neberie do úvahy. Keďže môj kamarát nedal konkrétnu konkurenčnú hypotézu, ktorá by dávala lepšie odhady, jeho téza musela byť, že štandardný evolučný model je vzhľadom na dané údaje *hlúpy* – že štandardný model dáva konkrétnu predpoveď, ktorá je nesprávna; že tomuto modelu sa darí horšie než úplnej nevedomosti alebo nejakej inej štandardnej konkurencii.

Najprv som padol do tejto pasce; prijal som implicitný predpoklad, že štandardný model predpovedá plynulosť a založil som svoj argument na spomienke, že zmeny vo fosílnom zázname neboli až také prudké, ako tvrdil on. Vyzval ma, aby som mu ukázal evolučný medzistupeň medzi *Homo erectus* a *Homo sapiens*; ja som googloval a našiel *Homo heidelbergensis*. Zablahoželal mi a uznal, že som získal významný bod, ale stále trval na tom, že tieto zmeny boli príliš ostré, a nie dosť rovnomerné. Začal som vysvetľovať, prečo si myslím, že zo štandardného modelu *mohol* vzniknúť vzor nerovnomernej zmeny: selekčné tlaky prostredia nemuseli byť stále rovnaké... „Aha!“ povedal môj kamarát, „chystáš si dopredu výhovorky.“

Predstavme si však, že by fosílny záznam namiesto toho ukazoval hladké a postupné zmeny. Mohol by môj kamarát argumentovať, že štandardný model evolúcie ako chaotického procesu plného šumu nemohol spôsobiť takúto hladkosť? Ak je vedeckým hriechom tvrdiť post facto, že naša obľúbená hypotéza predpovedá dané údaje, nemalo by byť takisto hriechom tvrdiť post facto, že konkurenčná hypotéza je na daných údajoch hlúpa?

Ak má hypotéza *čisto* technický model, nie je žiaden problém; môžeme spočítať predpoveď modelu formálne, bez neformálnych premenných, ktoré by poskytli možnosť na zasahovanie post facto. Ale čo s polotechnickými teóriami? Je zrejmé, že polotechnická teória musí *o niečom* dávať nejaké dobré predpovede vopred, lebo inak načo nám je? Ale *potom*, čo je teória polo-potvrdená, môžu oponenti tvrdiť, že údaje ukazujú problém v tejto polotechnickej teórii, keď je tento „problém“ vytvorený post facto? Prinajmenšom musia byť oponenti veľmi konkrétni ohľadom toho, ktoré údaje potvrdený model predpovedá hlúpo, a prečo tento potvrdený model musí urobiť (post facto) túto hlúpu predpoveď. Aká ostrá zmena je kvantitatívne „príliš ostrá“ na to, aby ju štandardný model evolúcie dovolil? Koľko presne rovnomernosti si myslíte, že štandardný model evolúcie predpovedá? Ako to viete? Je príliš neskoro to povedať, keď už ste videli dané údaje?

Keď ma môj kamarát obvinil z toho, že sa vyhováram, zastavil som sa a opýtal som sa sám seba, ktoré výhovorky očakávam, že budem potrebovať. Došiel som k záveru, že moje terajšie chápanie evolučnej teórie nehovorí nič o tom, či by miera evolučnej zmeny mala byť prerušovaná a zubatá, alebo hladká a postupná. Keby som daný graf nebol vopred videl, nedokázal by som ho predpovedať. (Nanešťastie, aj k tomuto záveru som došiel až potom, čo som videl tie údaje...) Možno v evolučnej rodine existujú modely, ktoré by vopred predpovedali rovnomernosť alebo premenlivosť, ale ak je to tak, ja ich nepoznám. A čo je podstatnejšie, môj kamarát ich tiež nepozná.

Nie je vždy múdre pýtať sa odporcov nejakej teórie, čo ich konkurencia predpovedá. Vypýtajte si predpovede teórie od jej najlepších zástancov. Akurát sa poistite, že ich predpovede zapíšete vopred. Áno, niekedy sa zástancovia teórie snažia, aby teória „zapadla“ do dôkazov, do ktorých jasne nezapadá. Ale ak zistíte, že sami neviete, čo nejaká teória predpovedá, opýtajte sa najprv jej zástancov, a potom požiadajte odporcov o krížovú kontrolu.

Navyše: Model môžu obsahovať šum. Ak máme hypotézu, že údaje majú sklon pomaly a trvalo stúpať, ale náš merací nástroj má chybu 5 %, potom nie je na nič dobré ukázať na dátový bod, ktorý klesne pod predchádzajúci dátový bod, triumfálne kričať: „Vidíte! Kleslo to! Dole, dole, dole! A nehovorte mi, prečo vaša teória zapadá do tohto poklesu; iba sa vyhovárate!“ Formálne, technické modely často obsahujú explicitné chybové členy. Chybový člen rozširuje hustotu podmienenej pravdepodobnosti, znižuje presnosť modelu, a zmenšuje skóre tejto teórie, ale stále platí bayesovské bodovacie pravidlo. Technický model môže pripustiť chyby, a robiť chyby, a stále byť lepší ako nevedomosť. V našom príklade so supermarketom, ešte aj tá presná hypotéza čísla 51 stále stavia iba 90 % svojej masy pravdepodobnosti na 51; presná hypotéza tvrdí iba, že 51 vyjde v deviatich prípadoch z desiatich. Ignorovať deväť 51-tiek, ukázať na jeden prípad 82, a triumfálne vykríknuť, to nie je vyvrátenie. Toto nie je výhovorka, je to explicitný predpoveď vopred technického modelu.

Chybový člen robí túto „presnú“ teóriu zraniteľnou voči superpresnej alternatíve, ktorá predpovedala 82. Štandardný model by bol zraniteľný aj voči presne nevedomému modelu, ktorý predpovedal šancu 60 % čísla 51 v tom kole, keď sme videli 82, a rozšíril tak podmienenú pravdepodobnosť pri tejto konkrétnej chybe entropickejšie. Bez ohľadu na to, aká dobrá je teória, veda má vždy miesto na lepšie bodujúcu konkurenciu. Ale ak *nepredložíte* lepšiu alternatívu, ak sa iba pokúšate ukázať, že prijatá teória je *hlúpa* vzhľadom na údaje, toto vedecké úsilie môže byť *náročnejšie* než iba nahradenie starej teórie novou.

Astronómovia zaznamenali nevysvetlený posun perihélia Merkúru, o ktorom newtonovská fyzika nevedela – alebo skôr, newtonovská fyzika predpovedala 5557 uhlových sekúnd za storočie, a pozorované množstvo bolo 5600.[232] Mali však vtedajší vedci zahodiť newtonovskú gravitáciu na základe takejto malej, nevysvetlenej kontraindície? Čo by používali namiesto nej? Nakoniec *bola* Newtonova gravitačná teória odsunutá bokom, keď Einsteinova všeobecná relativita presne vysvetlila nesúlad obežnej dráhy Merkúru a urobila aj úspešné predpovede vopred. Ale nedalo sa *dopredu* vedieť, že veci nakoniec dopadnú takto.

V devätnástom storočí bola trvalá anomália v obežnej dráhe Uránu. Ľudia hovorili: „Možno Newtonov zákon vo veľkých vzdialenostiach prestáva platiť.“ Nakoniec sa nejakí bystrí chlapci pozreli na túto anomáliu a povedali: „Mohla by to byť neznáma vonkajšia planéta?“ Urbain Le Verrier a John Couch Adams nezávisle niečo načmárali a zisťovali, pomocou Newtonovej štandardnej teórie – a predpovedali polohu Neptúna s presnosťou jedného uhlového stupňa, čím dramaticky *potvrdili* newtonovskú gravitáciu.[233]

Až *potom*, čo všeobecná relativita presne vrátila posun perihélia Merkúru, *vedeli* sme, že newtonovská gravitácia by to nikdy nevysvetlila.

* * *

V *Intuitívnom vysvetlení* sme videli, ako sa vhľad Karla Poppera, že falzifikácia je silnejšia než potvrdenie, prekladá do bayesovskej pravdy o pomeroch podmienených pravdepodobností. Popper sa mýlil, keď si myslel, že falzifikácia je *kvalitatívne iná* ako potvrdenie; obe sa riadia tými istými bayesovskými pravidlami. Popperova filozofia však odrazila dôležitú pravdu o kvantitatívnych rozdieloch medzi falzifikáciou a potvrdením.

> Na Poppera urobili hlboký dojem rozdiely medzi údajne „vedeckými“ teóriami Freuda a Adlera, a revolúciou spôsobenou Einsteinovou teóriou relativity vo fyzike v prvých dvoch desaťročiach tohto storočia. Hlavný rozdiel medzi nimi, ako to videl Popper, bol že kým Einsteinova teória bola vysoko „riskantná“, v tom zmysle, že z nej bolo možné vyvodzovať dôsledky, ktoré boli vo svetle vtedy dominantnej newtonovskej fyziky vysoko nepravdepodobné (napríklad že svetlo sa ohýba smerom k pevným telesám – čo potvrdil Eddingtonov experiment v roku 1919), a ktoré mohli, keby sa ukázali ako nepravdivé, falzifikovať celú teóriu, nič nemohlo, ani v princípe, falzifikovať psychoanalytické teórie. Tie

232 Kevin Brown, *Reflections On Relativity* (Raleigh, NC: printed by author, 2011), 405-414, http://www.mathpages.com/rr/rrtoc.htm.

233 Tamtiež.

> druhé, ako začal Popper cítiť, majú viac spoločné s primitívnymi poverami než so skutočnou vedou. Čo znamená, že uvidel, že to, čo je napohľad hlavným zdrojom sily psychoanalýzy, a základným princípom, na ktorom je založený jej nárok na vedecké postavenie, čiže jej schopnosť prijať a vysvetliť všetky možné formy ľudského správania, je vlastne kritická slabosť, pretože znamená, že nedokáže a nemôže dokázať naozaj predpovedať. Psychoanalytické teórie sú svojou podstatou nedostatočne presné na to, aby mali negatívne dôsledky, a tak sú odolné voči experimentálnej falzifikácii...
>
> Popper potom zavrhol indukciu, a odmietol názor, že je to charakteristická metóda vedeckého skúmania a odvodzovania, a na jej miesto dosadil falzifikovateľnosť. Je ľahké, tvrdí, získať indície v prospech prakticky hocijakej teórie, a v súlade s tým tvrdí, že takéto „potvrdenie“, ako to nazýva, by sa malo vedecky uznávať iba ak je pozitívnym výsledkom naozaj „riskantnej“ predpovede, ktorá sa ľahko mohla ukázať nepravdivou. Podľa Poppera je teória vedecká iba ak si vieme predstaviť udalosť, ktorá by ju vyvrátila. Každý skutočný test vedeckej teórie je potom logicky pokusom vyvrátiť ju, alebo falzifikovať ju...
>
> Každá skutočne vedecká teória je potom, podľa Poppera, zakazujúca, v tom zmysle, že z nej vyplýva zákaz určitých udalostí.[234]

Podľa Popperovej filozofie, sila vedeckej teórie nie je v tom, koľko vysvetľuje, ale koľko *nevysvetľuje*. Cnosť vedeckej teórie nespočíva vo výsledkoch, ktoré *pripúšťa*, ale vo výsledkoch, ktoré zakazuje. Freudove teórie, ktoré sa zdalo, že vysvetľujú všetko, *nezakazovali* nič.

Preložené do bayesovských pojmov, zisťujeme, že čím viac výsledkov nejaký model *zakazuje*, tým viac hustoty pravdepodobnosti tento model sústreďuje do zvyšných, povolených výsledkov. Čím viac výsledkov teória zakazuje, tým väčší je obsah poznania tejto teórie. Čím odvážnejšie sa teória vystavuje falzifikácii, tým jednoznačnejšie vám hovorí, aké vnemy máte očakávať.

Teória, ktorá dokáže vysvetliť *ľubovoľný* vnem zodpovedá hypotéze úplnej nevedomosti – uniformnej distribúcii s hustotou pravdepodobnosti rozloženou rovnomerne na všetky možné výsledky.

* * *

Flogiston bol odpoveďou 18. storočia na element ohňa gréckych alchymistov. Nemohli ste použiť teóriu flogistonu, aby ste predpovedali výsledok nejakej chemickej reakcie – najprv ste sa pozreli na výsledok, a potom ste použili flogiston, aby ste to vysvetlili. Teória flogistonu bola nekonečne pružná; maskovaná hypotéza nulového poznania. Podobne, teória *vitalizmu* nevysvetľuje, ako sa ruka hýbe, ani vám nehovorí, aké transformácie máte očakávať od organickej chémie; a určite nedovoľuje žiadne kvantitatívne výpočty.

Z opačnej strany:

Vyhnite sa kontrole podľa zoznamu: Mať *posvätné* tajomstvo alebo tajomnú odpoveď nie je to isté ako odmietať niečo vysvetliť. Niektoré prvky v našej fyzike berieme ako „základné“, zatiaľ nie ďalej redukované alebo vysvetlené. Ale tieto základné prvky našej fyziky sa riadia jasne definovanými, matematicky jednoduchými, formálne vypočítateľnými kauzálnymi pravidlami.

Občas nejaký šarlatán namieta voči modernej fyzike na základe toho, že neposkytuje „základný mechanizmus“ pre nejaký matematický zákon, ktorý sa momentálne považuje za základný. (Tvrdiť, že matematickému zákonu chýba „základný mechanizmus“ je jednou z položiek na *Indexe šarlatánov* Johna Baeza.[235]) Ten „základný mechanizmus“, ktorý navrhuje daný šarlatán ako odpoveď, je nejasný, slovný, a nijako nezvyšuje schopnosť predpovedať – inak by sme dotyčného neklasifikovali ako šarlatána.

Naša súčasná fyzika považuje elektromagnetické pole za základné, a odmieta ho ďalej vysvetliť. Ale toto „elektromagnetické pole“ je základ ovládaný jasnými matematickými pravidlami, a nemá žiadne vlastnosti mimo týchto matematických pravidiel, podlieha formálnemu výpočtu ktorý opisuje jeho kauzálny účinok na svet. Jedného dňa môže niekto navrhnúť lepšiu matematiku, ktorá dáva lepšie

234 Stephen Thornton, „Karl Popper,“ in *The Stanford Encyclopedia of Philosophy*, Winter 2002, ed. Edward N. Zalta (Stanford University), http://plato.stanford.edu/archives/win2002/entries/popper/.

235 John Baez, „The Crackpot Index,“ 1998, http://math.ucr.edu/home/baez/crackpot.html.

predpovede, ale neobviňoval by som terajší model z tajomnosti. Teória, ktorá zahŕňa *základné prvky* nie je to isté ako teória, ktorá obsahuje *tajomné prvky*.

Základy by mali byť jednoduché. „Život“ nie je dobrý základ; „kyslík“ je dobrý základ, a „elektromagnetické pole“ je ešte lepší základ. Život môže pripadať jednoduchý vitalistovi – je to jednoduchá, magická schopnosť vašich svalov pohybovať sa podľa príkazov vašich myšlienok. Prečo by život nemal byť vysvetlený jednoduchou, magickou látkou ako je *élan vital*? Lenže javy, ktoré vyzerajú *psychologicky* veľmi jednoducho – malé bodky svetla na oblohe, oranžovo jasné horúce plamene, mäso pohybujúce sa podľa príkazu myšlienok – často skrývajú rozhľahlé hlbiny základnej zložitosti. Tvrdenie, že život je zložitý jav môže vitalistovi pripadať neuveriteľne, keď pozerá na prázdne nepriesvitné tajomstvo bez viditeľných držadiel; lenže áno, Virginia, pod tým je základná zložitosť. Kritérium jednoduchosti, ktoré je relevantné pre Occamovu britvu, je *matematická* alebo *výpočtová* jednoduchosť. Až keď rozoberieme náš model na matematicky jednoduché základné prvky, ktoré samotné neobsahujú tajomné vlastnosti daného tajomstva, pôsobia na seba jasne definovanými spôsobmi a vytvárajú tak predtým tajomný jav ako podrobnú predpoveď, to je natoľko ne-tajomné, ako ľudstvo kedy zistilo, že sa niečo dá urobiť.

* * *

Mnoho ľudí na tomto svete verí, že keď zomrú, stretnú sa s prísne zazerajúcim chlapom meno Sv. Peter, ktorý preskúma činy v ich živote a vypočíta im skóre za morálku. Predpokladá sa, že bodovacie pravidlo Sv. Petra je jedinečné a nemení sa pri triviálnych zmenách uhla pohľadu. Nanešťastie veriaci nedokážu získať kvantitatívnu, presne vypočítateľnú špecifikáciu tohto bodovacieho pravidla, čo vyzerá dosť neférovo.

Náboženstvo *bayesianizmus* tvrdí, že váš večný osud závisí od pravdepodobnostných úsudkov, ktoré ste urobili v živote. Na rozdiel od primitívnejších vier, bayesianizmus dokáže dať kvantitatívnu, presne vypočítateľnú špecifikáciu, ako bude určený váš večný osud.

Naše správne bayesovské bodovacie pravidlo poskytuje spôsob, ako spočítať skóre pre viaceré experimenty, a toto skóre sa nemení podľa toto, ako rozdelíme tieto „experimenty“ alebo v akom poradí pozbierame výsledky. Sčítavame logaritmy pravdepodobností. Toto zodpovedá ich vynásobeniu pravdepodobností pripísaných výsledku každého experimentu, aby sme zistili spoločnú pravdepodobnosť všetkých experimentov dokopy. Používame logaritmus, aby sme zjednodušili naše intuitívne chápanie nazbieraného skóre, aby sme si udržali kontrolu nad malými zlomkami, aby sme zabezpečili, že maximalizujeme naše *očakávané* skóre pomocou vyslovenia svojich úprimných pravdepodobností namiesto stavenia všetkých svojich hracích peňazí na tú najpravdepodobnejšiu stávku.

Bayesianizmus hovorí, že keď zomriete, Pierre-Simon Laplace preskúma každú jednu udalosť vo vašom živote, od ranného nájdenia svojich topánok vedľa postele, po nájdenie svojho pracoviska na jeho zvyčajnom mieste. Každý prehrávajúci žreb v lotérii znamená, že vám záležalo natoľko, aby ste hrali. Laplace odhadne pravdepodobnosť, ktorú ste vopred priradili každej udalosti. Kde ste vopred nepriradili presnú číselnú pravdepodobnosť, Laplace preskúma váš stupeň očakávania prekvapenia, extrapoluje ostatné možné výsledky a vaše extrapolované reakcie, a renormalizuje vaše extrapolované emócie na distribúciu podmienenej pravdepodobnosti pre možné výsledky. (Odtiaľ pojem „Laplaceovská superinteligencia“.)

Potom Laplace vezme všetky udalosti vo vašom živote, a každú pravdepodobnosť, ktorú ste priradili každej udalosti, a vynásobí všetky tieto pravdepodobnosti dokopy. Toto je váš Posledný Súd – pravdepodobnosť, ktorú ste priradili svojmu životu.

Tí, ktorí sa riadia bayesianizmom sa usilujú celý svoj život maximalizovať svoj Posledný Súd. Toto je jediná cnosť bayesianizmu. Zvyšok je iba matematika.

Pamätajte: cesta bayesianizmu je prísna. Akú pravdepodobnosť prisúdite každé ráno výroku: „Vyjde slnko?“ (Nebudeme počítať také slovíčkarenie ako zamračené dni, alebo že Zem obieha okolo Slnka.) Možno niekto, kto sa neriadi bayesianizmom, by bol pokorný a dal pravdepodobnosť 99,9 %. Ale my, ktorí sa riadime bayesianizmom, odhodíme všetky ohľady na skromnosť a drzosť, a budeme sa usilovať iba maximalizovať svoj Posledný Súd. Ako posadnutý hráč počítačových hier, záleží nám iba na

tomto číselnom skóre. Budeme sa s touto témou východu Slnka stretávať 365-krát ročne, takže by sme mohli výrazne zlepšiť svoj Posledný Súd, ak upravíme svoje priradenie pravdepodobnosti.

Ako to je, dokonca aj keby Slnko vyšlo každé ráno, každý rok sa náš Posledný Súd zníži o činiteľ $0{,}999^{356} = 0{,}7$, zhruba -0,52 bitov. Každé dva roky sa náš Posledný Súd zníži o viac než keby sme zistili, že nepoznáme výsledok hodenia mincou! To nemožno tolerovať. Ak zvýšime svoju dennú pravdepodobnosť východu slnka na 99,99 %, potom sa každý rok náš Posledný Súd zníži iba o činiteľ 0,964. Lepšie. Ale aj tak, v tom nepravdepodobnom prípade, že budeme žiť presne 70 rokov a potom zomrieme, náš Posledný Súd bude iba 7,75 % toho, čo mohol byť. Čo ak priradíme východu slnka pravdepodobnosť 99,999 %? Potom bude o 70 rokov náš Posledný Súd vynásobený 77,4 %.

Prečo nepriradiť hodnotu 1,0?

Ten, kto sa riadi bayesianizmom, *nikdy* nepriradí hodnotu 1,0 *ničomu*. Priradiť nejakému výsledku pravdepodobnosť 1,0 spotrebuje *všetku* vašu masu pravdepodobnosti. Ak nejakému výsledku priradíte masu pravdepodobnosti 1,0, a skutočnosť doručí inú odpoveď, museli ste *skutočnému* výsledku priradiť pravdepodobnosť *nula*. Toto je jediný smrteľý hriech v bayesianizme. Nula krát hocičo je nula. Keď Laplace vynásobí všetky pravdepodobnosti vo vašom živote, výsledná pravdepodobnosť bude nula. Váš Posledný Súd bude úplné fiasko, zilch, nada, nil. Bez ohľadu na to, aké rozumné boli vaše odhady počas zvyšku vášho života, strávite večnosť vedľa nejakého chlapíka, ktorý veril v lietajúce taniere a čerpal všetky svoje informácie z Weekly World News. Opäť je užitočné použiť logaritmus a odhaliť nevinne znejúcu „nulu“ v jej skutočnej podobe. Riskovať pravdepodobnosť výsledku nula je ako prijať stávku, ktorej výplata je mínus nekonečno.

Čo ak sa ľudstvo rozhodne rozobrať hmotu Slnka na kúsky (hviezdne inžinierstvo), alebo vypnúť Slnko, lebo plytvá entropiou? Nuž, hovoríte, že to budete očakávať, že budete mať príležitosť zmeniť svoje priradenie pravdepodobnosti skôr než sa to naozaj strane. Čo ak sa u niekoho v pivnici umelá inteligencia rekurzívne sebazdokonalí na superinteligenciu, potajomky vyvinie nanotechnológiu, a jedného rána *ona* rozoberie Slnko? Ak počas poslednej noci na svete priradíte zajtrajšiemu východu slnka pravdepodobnosť 99,999 %, váš Posledný Súd klesne o činiteľ 100 000. Mínus 50 decibelov! Hrozné, však?

Aká je teda vaša najlepšia stratégia? No, predpokladajme, že očakávate na 50 %, že UI superinteligencia postavená v garáži rozoberie Slnko niekedy počas nasledujúcich 10 rokov, a potom odhadnete, že je zhruba rovnaká šanca, že sa toto stane v ktorýkoľvek deň medzi teraz a vtedy. Počas ľubovoľnej noci by ste očakávali na 99,98 %, že zajtra vyjde slnko. Ak je toto naozaj to, čo očakávate, potom nemáte motív povedať ako vašu pravdepodobnosť niečo iné ako 99,98 %. Ak sa cítite nervózne, že toto očakávanie je príliš nízke alebo príliš vysoké, nemôže to byť to, čo očakávate, keď zohľadníte túto nervozitu.

Ale hlbšia pravda bayesianizmu je toto: nemôžete prekabátiť systém. Nemôžete dať skromnú odpoveď, ani sebavedomú. Musíte zistiť presne nakoľko očakávate, že Slnko zajtra vyjde, a povedať toto číslo. Musíte oholiť každý chlp skromnosti alebo drzosti, a opýtať sa, či očakávate, že budete bodovaní podľa východu Slnka, alebo podľa toho, že nevyjde. Nepozerajte sa na svoje výhovorky, ale pýtajte sa, ktoré výhovorky očakávate, že budete potrebovať. Keď prídete na presný stupeň vášho očakávania, jediný spôsob, ako ďalej zvýšiť vás Posledný Súd je zlepšiť presnosť, kalibráciu, a rozlišovanie vášho očakávania. Nemôžete dopadnúť lepšie iným spôsobom ako lepšie hádať a presnejšie očakávať.

Ehm, vlastne, jedine keby ste spáchali samovraždu vo veku päť rokov, čím by se zabránili, aby váš Posledný Súd ďalej klesal. Alebo, keby sme do funkcie úžitku pridali ako záplatu nový hriech, nariadenie proti samovražde, mohli by ste utekať pred tajomstvom, vyhýbať sa všetkým situáciám, v ktorých si myslíte, že by ste mohli niečo nevedieť. Toľko k tomuto náboženstvu.

* * *

Ideálne by sme výsledok experimentu mali predpovedať vopred, pomocou nášho modelu, a potom môžeme vykonať experiment, aby sme videli, či je výsledok v súlade s naším modelom. Nanešťastie nemáme prúd informácií vždy pod kontrolou. Niekedy na nás Príroda hádže skúsenosti, a kým pomyslíme

na vysvetlenie, už sme videli tie údaje, ktoré máme vysvetliť. Toto bol jeden z vedeckých hriechov, ktorých sa dopustil evolucionizmus 19. storočia; Darwin pozoroval podobnosť mnohých druhov, a ich prispôsobenie miestnemu prostrediu, skôr než mu napadla hypotéza prirodzeného výberu. Evolucionizmus 19. storočia prišiel na svet ako vysvetlenie post facto, nie ako predpoveď vopred.

A toto nie je problém iba polotechnických teórií. V roku 1846 sa úspešná dedukcia existencie Neptúna na základe gravitačných odchýliek v obežnej dráhe Uránu považovala za veľký triumf Newtonovej teórie gravitácie. Prečo? Pretože existencia Neptúna bola prvým pozorovaním, ktoré potvrdilo predpoveď *vopred* o gravitácii Neptúna. Všetky ostatné javy, ktoré Newton vysvetlil, ako obežné dráhy a odchýlky obežných dráh a príliv, boli pozorované veľmi podrobne už predtým ako ich Newton vysvetlil. Nikto vážne nepochyboval, že Newtonova teória je správna. Newtonova teória vysvetľovala príliš veľa príliš presne, a nahradila zbierku ad-hoc modelov jedným jednotným matematickým zákonom. Napriek tomu sa predpoveď vopred existencie Neptúna, nasledovaná pozorovaním Neptúna na takmer presne predpovedanom mieste, považovala za prvý veľký triumf Newtonovej teórie v predpovedaní toho, čo žiaden predchádzajúci model nedokázal predpovedať. Medzi všeobecným prijatím Newtonovej teórie a prvou pôsobivou predpoveďou *vopred* newtonovskej gravitácie uplynul značný čas. V dobe, keď Newton prišiel so svojou teóriou, vedci už pozorovali, veľmi podrobne, väčšinu javov, ktoré newtonovská gravitácia predpovedala.

Lenže pravidlo predpovede vopred je morálka vedy, nie zákon teórie pravdepodobnosti. Ak ste už videli údaje, ktoré musíte vysvetliť, potom vás Veda môže zatratiť, ale vaše nepríjemnosť nezrúti zákony teórie pravdepodobnosti. Stane sa akurát to, že pre nešťastného človeka bude omnoho zložitejšie *dodržiavať* zákony teórie pravdepodobnosti. Keď sa rozhodujete, ako odhodnotiť hypotézu podľa bayesovského bodovacieho pravidla, potrebujete zistiť, koľko masy pravdepodobnosti priraďuje táto hypotéza pozorovanému výsledku. Ak musíme urobiť naše predpovede vopred, potom je ľahšie všimnúť si, keď sa niekto snaží tvrdiť, že každý možný výsledok je jeho predpoveďou vopred, keď používa priveľa masy pravdepodobnosti, keď je úmyselne nejasný, aby sa vyhol falzifikácii, a tak ďalej.

Žiaden numerológ nedokáže predpovedať, aké čísla vyhrajú v lotérii budúci týždeň, ale ochotne vám vysvetlí mystický význam čísel, ktoré vyhrali v lotérii minulý týždeň. Povedzme, že v lotérii Mega Ball minulý týždeň vyhralo čísla sedem, spomedzi 52 možných výsledkov. To sa samozrejme stalo preto, lebo sedem je šťastné číslo. Takže v lotérii Mega Ball budúci týždeň tiež vyhrá sedem? Samozrejme chápeme, že to nie je isté, ale ak je to šťastné číslo, mali by ste mu priradiť pravdepodobnosť vyššiu než 1/52... a potom budeme hodnotiť vaše odhady počas niekoľkých rokov, a ak je vaše skóre príliš nízke, tak vás zbičujeme... čo ste to povedali? Chcete priradiť pravdepodobnosť presne 1/52? Ale to je rovnaká pravdepodobnosť pre každé číslo; čo sa stalo s tým, že sedmička je šťastná? Nie, prepáčte, nemôžete prideliť pravdepodobnosť 90 % číslu sedem, a zároveň pravdepodobnosť 90 % číslu jedenásť. Rozumieme, že sú to obe šťastné čísla. Áno, rozumieme, že sú to *veľmi* šťastné čísla. Ale takto to nefunguje.

Dokonca aj keď poslucháč nevie o Bayesovej ceste a nepýta si formálne pravdepodobnosti, pravdepodobne bude podozrievať, ak sa snažíte pokryť príliš veľa možností. Predstavte si, že vás požiadajú, aby ste predpovedali, čo vyhrá budúci týždeň v Mega Ball, a vy pomocou numerológie vysvetlíte, prečo by lopta číslo jeden veľmi dobre zapadala do vašej teórie, a prečo by lopta číslo dva veľmi dobre zapadala do vašej teórie, a prečo by lopta číslo tri dobre zapadala do vašej teórie... aj ten najdôverčivejší poslucháč by mohol začať klásť otázky, keď by ste sa dostali po číslo dvanásť. Azda by ste nám mohli povedať, ktoré čísla sú nešťastné a jednoznačne v lotérii nevyhrajú? No, trinástka je nešťastná, ale nie je to absolútne *nemožné* (poisťujete sa, *očakávate* vopred, ktorú výhovorku možno budete potrebovať).

Ale ak vás požiadame, aby ste vysvetlili čísla v lotérii *minulého týždňa*, nuž, sedmička bola prakticky nevyhnutná. Táto sedmička by sa mala jednoznačne počítať ako veľký úspech modelu „šťastných čísel“ lotérie. A nemohla to ani náhodou byť trinástka; teória šťastia to jasne vylučuje.

* * *

Predstavte si, že sa jedného dňa zobudíte a vaše ľavé rameno bolo nahradené modrým chápadlom. Toto modré chápadlo poslúcha vaše pohybové príkazy - môžete ho použiť na dvíhanie pohárov, riadenie auta, atď. Ako by ste vysvetlili tento hypotetický scenár? Zastavte sa na chvíľu a rozmýšľajte nad touto záhadou, než budete pokračovať.

(Priestor na rozmýšľanie...)

Ako by som ja vysvetliť udalosť, že moje ľavé rameno bolo nahradené modrým chápadlom? Odpoveď je, že by som to nevysvetľoval. Nestane sa to.

Bolo by dosť ľahké vytvoriť slovné vysvetlenie, ktoré „zapadá“ do hypotézy. Existuje veľa vysvetlení, ktoré môžu „zapadať“ do hocičoho, vrátane (ako špeciálny prípad „hocičoho“) nahradenia môjho ramena modrým chápadlom. Božský zásah je dobrým všeobecne použiteľným vysvetlením. Alebo mimozemšťania s ľubovoľnými motívmi a schopnosťami. Alebo som mohol byť šialený, halucinovať, snívať celý svoj život v nemocnici. Takéto vysvetlenia „zapadajú“ do všetkých výsledkov rovnako dobre, a rovnako slabo, rovnajú sa hypotéze úplnej nevedomosti.

Testom, či nejaký model skutočnosti „vysvetľuje“, že sa moje rameno premenilo na modré chápadlo, je či tento model sústreďuje významnú masu pravdepodobnosti do tohto *konkrétneho* výsledku. Prečo ten sen, v nemocnici? Prečo by mi mimozemšťania urobili konkrétne toto, na rozdiel od ďalšej miliardy vecí, ktoréby mohli urobiť? Prečo by sa moje rameno v toto ráno premenilo na chápadlo, keď zostalo ramenom každé iné ráno v mojom živote? A vo všetkých prípadoch musím hľadať argument dostatočne pádny na to, aby som túto konkrétnu predpoveď urobil *vopred*, nie púhu zlučiteľnosť. Keď už raz viem výsledok, je omnoho zložitejšie preosievať hypotézy, aby som našiel dobré vysvetlenia. Akúkoľvek hypotézu skúsim, bude pre mňa veľmi ťažké nepriradiť viac masy pravdepodobnosti včerajšiemu výsledku modrého chápadla, než keby som extrapoloval naslepo, hľadajúc *najpravdepodobnejšiu* predpoveď daného modelu pre zajtrajšok.

Model nie vždy predpovedá všetky vlastnosti údajov. Príroda nemá výhradný sklon predkladať mi riešiteľné problémy. Možno sa so mnou pohráva nejaké božstvo, a myseľ tohto božstva je výpočtovo nezvládnuteľná. Keď hodím vyváženou mincou, neexistuje spôsob, ako ďalej vysvetliť výsledok, žiaden model, ktorý by dával lepšie predpovede ako hypotéza maximálnej entropie. Ale ak predpokladám model, ktorý nemá žiadne vnútorné podrobnosti, alebo model, ktorý nerobí žiadne ďalšie predpovede, nielenže nemám dôvod veriť tomuto predpokladu, ale ani nemám dôvod sa oň zaujímať. Minulú noc bolo moje rameno nahradené modrým chápadlom. Prečo? Mimozemšťania! Čo teda urobia zajtra? Podobne, ak pripisujem toto modré chápadlo halucinácii, ako snívam svoj život ležiac v kóme, stále neviem nič viac o tom, čo budem halucinovať zajtra. Prečo by ma teda zaujímalo, či to boli mimozemšťania alebo halucinácia?

Čo by teda mohlo byť *dobrým* vysvetlením, keby som sa jedného rána zobudil a zistil, že sa moje rameno premenilo na modré chápadlo? Tvrdiť, že je niečo „dobrým vysvetlením“ tohto hypotetického zážitku, by si vyžadovalo taký argument, že keby som *teraz* uvažoval o tomto hypotetickom argumente, ešte *predtým* než sa moje rameno premenilo na modré chápadlo, šiel by som spať s obavami, že sa moje rameno *naozaj premení* na chápadlo.

Ľudia sa hrajú s prijateľnosťou, vysvetľujú udalosti, o ktorých neočakávajú, že sa niekedy naozaj stanú, hoci toto nevyhnutne porušuje zákony teórie pravdepodobnosti. Koľko ľudí, ktorí si myslelo, že by dokázali „vysvetliť“ hypotetický zážitok prebudenia sa s ramenom nahradeným chápadlom, by išlo spať s obavou, že sa im to naozaj môže stať? Keby mali takú odvahu ako ich presvedčenia, povedali by: neočakávam, že sa niekedy stretnem s touto hypotetickou skúsenosťou, a preto ju nedokážem vysvetliť, ani nemám motív to skúšať. Takéto veci sa stávajú iba vo webových komixoch, a ja si nepotrebujem

pripravovať vysvetlenia, pretože v skutočnom živote nikdy nebudem mať šancu ich použiť. Ak sa niekedy ocitnem v tejto nemožnej situácii, nechcem prísť ani o bodku svojho cenného zmätku.

Pre bayesovca sú pravdepodobnosti očakávania, nie púhe názory, ktoré vyhlasujeme zo striech. Ak mám model, ktorý priraďuje masu pravdepodobnosti môjmu prebudeniu sa s modrým chápadlom, potom som nervózny, že sa možno zobudím s modrým chápadlom. Čo ak je tento model fantastický, napríklad, že bosorka začaruje kúzlo, ktoré ma premiestni do náhodne vybraného webového komixu? Potom je *pôvodná pravdepodobnosť* webkomixového bosoráctva taká nízka, že moje chápanie *skutočného sveta* tejto hypotéze nepripisuje žiadnu významnú váhu. Hypotéza bosoráctva, keby sme ju brali ako danú, by mohla priradiť nezanedbateľnú podmienenú pravdepodobnosť prebudeniu sa s modrým chápadlom. Ale moje očakávanie tejto hypotézy je také nízke, že neočakávam žiadne predpovede tejto hypotézy. To, že si dokážem predstaviť hypotézu bosoráctva, by nijako nemalo zmenšiť môj totálny zmätok, keby som sa naozaj zobudil s chápadlom, pretože pravdepodobnosť, ktorú v skutočnom svete prisudzujem hypotéze bosoráctva je prakticky nula. Moja hypotéza s nulovou pravdepodobnosťou by mi nepomohla *vysvetliť*, prečo som sa zobudil s chápadlom, pretože tento argument nie je dosť dobrý na to, aby som teraz *očakával*, že sa zobudím s chápadlom.

V zákonoch teórie pravdepodobnosti sú distribúcie podmienenej pravdepodobnosti pevne danými vlastnosťami nejakej hypotézy. V umení rozumnosti, *vysvetliť* znamená *očakávať*. *Očakávať* znamená *vysvetliť*. Predstavte si, že som lekársky výskumník a v bežnom procese realizácie svojho výskumu si všimnem, že moja chytrá nová teória anatómie zdanlivo umožňuje malú a nejasnú možnosť, že sa moje rameno premení na modré chápadlo. „Haha!“ poviem, „aké zvláštne a bláznivé!“ a budem sa cítiť trošku nervózne. *Toto* by bolo dobré vysvetlenie pre zobudenie sa s chápadlom, keby sa to niekedy stalo.

Ak ma reťaz úvah nerobí nervóznym, vopred, ohľadom zobudenia sa s chápadlom, potom by toto uvažovanie bolo slabým vysvetlením, keby sa daná udalosť *stala*, pretože kombinácia pôvodnej pravdepodobnosti a podmienenej pravdepodobnosti bola príliš nízka na to, aby som tomuto výsledku priradil nejakú významnú pravdepodobnosť v skutočnom svete.

Ak začnete od dobre kalibrovaných pôvodných predpokladov a použijete bayesovské uvažovanie, skončíte s dobre kalibrovanými závermi. Predstavte si, že dva milióny bytostí, roztrúsených po rôznych planétach vo vesmíre, majú príležitosť stretnúť sa s niečím takým zvláštnym ako je prebudiť sa s chápadlom (alebo – ach! - s desiatimi prstami). Jeden milión z týchto bytostí hovorí „jedna z tisíca“ ako pôvodnú pravdepodobnosť nejakej hypotézy X, a každá hypotéza X hovorí „jedna zo sto“ ako podmienenú pravdepodobnosť, že sa zobudí s chápadlom. A jeden milión z týchto bytostí hovorí „jedna zo sto“ ako pôvodnú pravdepodobnosť nejakej hypotézy Y, a každá hypotéza Y hovorí „jedna z desať“ ako podmienenú pravdepodobnosť prebudenia sa s chápadlom. Ak predpokladáme, že všetky tieto bytosti sú dobre kalibrované, potom by sme sa mali pozrieť po vesmíre a nájsť desať bytostí, ktoré skončia s chápadlom kvôli hypotézam z triedy dôveryhodnosti X, a tisíc bytostí, ktoré skončia s chápadlom kvôli hypotézam z triedy dôveryhodnosti Y. Takže ak zistíte, že máte chápadlo, a *ak* sú vaše pravdepodobnosti dobre kalibrované, potom je pravdepodobnejšie, že toto chápadlo pochádza z hypotézy, ktorú by ste zaradili ako pravdepodobnú, než z hypotézy, ktorú by ste zaradili ako nepravdepodobnú. (Čo ak sú vaše pravdepodobnosti tak slabo kalibrované, že keď poviete „milión k jednej“, stane sa to v jednom prípade z dvadsiatich? Potom máte prehnanú sebadôveru a my prispôsobíme vaše pravdepodobnosti smerom k menšiemu rozlíšeniu / väčšej entropii.)

Hypotéza, že ste boli premiestnení do webového komixu, dokonca aj keď „vysvetľuje“ scenár, že sa zobudíte s modrým chápadlom, je slabým vysvetlením, pretože má nízku pôvodnú pravdepodobnosť. Hypotéza webového komixu neprispieva k vysvetleniu chápadla, pretože nespôsobuje, že očakávate, že sa zobudíte s chápadlom.

Ak začneme s biliardou vedomých myslí roztrúsených po vesmíre, dosť veľa z týchto bytostí zažije udalosti, ktoré sú veľmi pravdepodobné, asi len milión bytostí zažije udalosti, ktorých pravdepodobnosť za celý život je miliarda k jednej (ako by sme očakávali, sledujúc nespočetnými očami a s dokonalou kalibráciou), a ani jedna bytosť nezažije nič nemožné.

Ak by ste sa nejako zobudili s chápadlom, bolo by to pravdepodobne kvôli niečomu omnoho pravdepodobnejšiemu než „premiestnenie do webového komixu“, nejaký dokonale normálny dôvod prebudenia sa s chápadlom, ktorý ste skrátka nečakali. Aký napríklad dôvod? Neviem. Nič. Ja neočakávam, že sa zobudím s chápadlom, takže vám na to neviem dať žiadnu dobrú odpoveď. Prečo by som sa unúval zostavovať výhorovky, keď nečakám, že ich použijem? Keby som sa obával, že jedného dňa možno budem potrebovať chytrú výhovorku, prečo som sa zobudil s chápadlom, *dôvod, prečo som z tejto možnosti nervózny*, by bol *mojím* vysvetlením.

Skutočnosť rozdáva zážitky pomocou pravdepodobnosti, nie dôveryhodnosti. Ak niekedy zistíte, že váš laptop porušuje zachovanie hybnosti, potom si skutočnosť musí myslieť, že je dokonale normálne, aby vám to urobila. Ako by porušenie zachovania hybnosti mohlo byť dokonale normálne? Očakávam, že táto otázka nemá odpoveď, a že ju nebude treba zodpovedať. Podobne sa ľudia *nebudia* s chápadlami, takže to zrejme *nie je* dokonale normálne.

* * *

Existuje zdrvujúca pravda, taká prekvapivá a desivá, že sa ľudia celou svojou silou bránia jej dôsledkom. Napriek tomu existuje pár osamelých jedincov s odvahou prijať toto satori. Tu je tá múdrosť, ak ste jedným z múdrych:

Od samého začiatku

Sa nikdy nestala

Jediná nezvyčajná vec

Beda tým, ktorí odvracajú svoje oči od zebier a snívajú o drakoch! Ak sa nedokážeme naučiť mať radosť z púhej skutočnosti, naše životy budú veru prázdne.

*→

—

→ http://yudkowsky.net/rational/technical/

Kniha V.

Obyčajné dobro

Ciele: Úvod

(napísal Rob Bensinger)

Teória hodnoty je štúdium toho, na čom ľuďom záleží. Je to štúdium našich cieľov, našich záľub, našich potešení a trápení, našich strachov a našich ambícií.

Zahŕňa aj konvenčnú morálku. Teória hodnoty zahŕňa aj veci, o ktorých by sme *chceli*, aby nám na nich záležalo, alebo na ktorých by nám záležalo, keby sme boli múdrejšími a lepšími ľuďmi -– nie iba veci, na ktorých nám záleží teraz.

Teória hodnoty zahŕňa aj všedné, každodenné hodnoty: umenie, jedlo, sex, priateľstvo, a všetko ostatné, čo dáva životu jeho citový náboj. Ísť do kina s vaším kamarátom Samom môže byť niečo, čo si ceníte, hoci to nie je *morálna* hodnota.

Považujeme za užitočné všímať si svoje vlastné hodnoty a debatovať o nich, pretože to, ako sa správame, nie je vždy to, ako by sme sa správať chceli. Naše preferencie môžu byť vo vzájomnom konflikte. Môžeme túžiť, aby sme túžili po iných veciach. Môže nám chýbať vôľa, pozornosť, alebo vhľad potrebné na to, aby sme konali tak, ako by sme radi konali.

Ľudia sa zaujímajú o dôsledk svojich činov, ale nie dostatočne konzistentne, aby sa formálne kvalifikovali ako činitelia s funkciami úžitku. To, že ľudia nekonajú tak, ako by konať chceli, označujeme slovami: „ľudia nie sú inštrumentálne rozumní."

Teória a prax

Aby to bolo ešte zložitejšie, existuje priepasť medzi tým, ako si *myslíme*, že by sme chceli konať, a ako by sme *naozaj* chceli konať.

Filozofi – a psychológovia, a politici – si vášnivo odporujú v názoroch na to, čo chceme, a čo by sme mali chcieť. Odporujú si dokonca aj v tom, *čo to znamená*, že by človek „mal" niečo chcieť. Dejiny morálnej teórie, a dejiny ľudských snáh o koordináciu, sú dláždené mŕtvolami neúspešných Vodiacich Princípov k Skutočnej Dokonalej Nie-Teraz-To-Myslím-Naozaj Normativite.

Ak sa snažíte vymyslieť *spoľahlivú* a *pragmaticky užitočnú* špecifikáciu svojich cieľov – nie iba kvôli vyhrávaniu filozofických debát, ale (napríklad) kvôli dizajnovaniu bezpečnej autonómnej prispôsobivej UI, alebo kvôli vybudovaniu funkčných inštitúcií a organizácií, alebo aby ste sa ľahšie rozhodli, ktorej charite darovať, alebo aby ste zistili, ktoré cnosti by ste si mali rozvíjať – doterajšie úspechy ľudstva v teórii hodnoty vám nedávajú veľkú nádej.

Obyčajné dobro zahŕňa tri postupnosti článkov z blogu o ľudských hodnotách: „Falošné preferencie" (o neúspešných pokusoch o teóriu hodnoty), „Teória hodnoty" (o prekážkach pri vyvíjaní novej teórie, a niektoré intuitívne žiadúce vlastnosti takejto teórie), a „Kvantitatívny humanizmus" (o zákernej otázke, ako by sme takéto teórie mali *aplikovať* na svoje bežné morálne intuície a rozhodovanie).

Tá posledná z týchto tém je najdôležitejšia. Normatívna teória je užitočná natoľko, nakoľko dobre sa dá preložiť do morálnej praxe. Získať hlbšie a plnšie pochopenie vašich hodnôt by vám malo pomôcť pri ich skutočnom napĺňaní. Ako absolútne minimum, vaša teória by vám *nemala prekážať* pri vašej praxi. Načo by inak bolo dobré vedieť, čo je dobré?

Zjednocovanie tohto umenia aplikovanej etiky (a aplikovanej estetiky, a aplikovanej ekonómie, a aplikovanej psychológie) s našimi najlepšími dostupnými údajmi a teóriami často končí pri otázke, kedy by sme mali dôverovať našim okamžitým úsudkom, a kedy by sme sa ich mali zbaviť.

V mnohých prípadoch sú naše explicitné modely toho, na čom nám záleží, také hmlisté alebo nepraktické, že sme na tom lepšie, ak dôverujeme svojim nejasným prvým dojmom. V mnohých iných prípadoch *dokážeme* robiť veci lepšie s informovanejším a systematickým prístupom. Neexistuje jedna odpoveď na všetko. Budeme musieť skúmať príklady a skúsiť si všimnúť rôzne varovné znaky pre „tu zvyknú zlyhávať sofistikované teórie" a „tu zvyknú zlyhávať naivné pocity".

Cesta a cieľ

Opakovaná téma na nasledujúcich stranách je otázka: *Kam pôjdeme? Ktoré výsledky sú naozaj hodnotné?*

V odpoveď na túto otázku Yudkowsky vymyslel pojem „teória zábavy“. Teória zábavy je pokus zistiť, ako by mohla vyzerať naša ideálna predstava budúcnosti – nie iba systém vlády alebo morálny kódex, pod ktorými by sme žili, ale dobrodružstvá, do ktorých by sme sa ideálne púšťali, hudba, ktorú by sme ideálne skladali, a všetko ostatné, čo v konečnom dôsledku v živote chceme.

Z hľadiska budúcnosti sa otázky teórie zábavy pretínajú s otázkami *transhumanizmu*, názoru, že môžeme ľudskú situáciu radikálne zlepšiť, ak urobíme dostatočný vedecký a spoločenský pokrok.[236] Transhumanizmus súvisí s mnohými debatami v morálnej filozofii, napríklad či by najlepšie dlhodobé výsledky pre vedomý život boli založené na *hedonizme* (hľadaní potešenia) alebo na zložitejších pojmoch *eudaimonie* (všeobecného dobrého života). Ďalšie futuristické myšlienky diskutované v rôznych bodoch knihy *Rozumnosť: od algoritmov po zombie* sú *kryonika* (zmrazenie tela po smrti pre prípad, že budúca lekárska technológia nájde spôsob, ako vás oživiť), *nahrávanie mysle* (implementovanie ľudskej mysle v syntetickom hardvéri), a veľkorozmerná kolonizácia vesmíru.

Možno prekvapujúco, teória zábavy je jednou zo zanedbávanejších aplikácií teória hodnoty. Plánovanie utópie je pomerne pasé – čiastočne preto, lebo zaváňa naivnosťou, a čiastočne preto, lebo skúsenosť ukazuje, že s premieňaním utópií na skutočnosť sme na tom *biedne*. Dokonca aj samotné slovo „utópia“ odráža tento cynizmus; pochádza z gréckeho „nie miesto“.

Ak sa však vzdáme hľadania skutočnej, dosiahnuteľnej utópie (nazvime ju *eutópia*, „dobré miesto“), nie je zrejmé, že kumulatívny účinok nášho dosahovania krátkodobých cieľov bude budúcnosť, ktorú by sme dlhodobo považovali za hodnotnú. Hodnota nie je nevyhnutnou súčasťou sveta. Vytvoriť ju je práca. Zachovať ju je práca.

To vyvoláva druhú otázku: *Ako sa tam dostaneme? Aký je vzťah medzi dobrými cieľmi a dobrými prostriedkami?*

Keď hráme nejakú hru, chceme si vychutnať samotný proces. Vo všeobecnosti netúžime to celé preskočiť a byť vyhlásení za víťazov. Niekedy je cesta dôležitejšia než cieľ. Niekedy je cesta to *jediné* dôležité.

Ale existujú aj prípady, keď je to presne naopak. Niekedy je výsledný stav príliš dôležitý na to, aby sme „cestu“ zohľadňovali vo svojich rozhodnutiach. Ak sa snažíte zachrániť život svojho príbuzného, nie je nevyhnutne *zlé* užiť si niečo z toho procesu; ale ak dokážete výrazne zvýšiť svoju pravdepodobnosť úspechu tým, že si vyberiete menej zábavnú stratégiu...

V mnohých prípadoch sa naše hodnoty sústredia vo výsledkoch našich činov, a v našej budúcnosti. Záleží nám na tom, ako svet dopadne – najmä tie časti sveta, ktoré dokážu prežívať lásku, bolesť, túžbu.

Ako sa môžu neosobné, abstraktné teórie v takýchto prípadoch porovnávať so živými emóciami? Ešte všeobecnejšie: Aký je morálny vzťah medzi činmi a dôsledkami?

Toto sú ťažké otázky, ale možno dokážeme prinajmenšom urobiť pokrok v pochopení, čo nimi *myslíme*. Na čom zakladáme našu predstavu o tom, čo je „hodnotné“ na samotnom začiatku nášho vypytovania?

---

236 Jedným príkladom transhumanistického argumentu je: „Mohli by sme reálne zrušiť starnutie a choroby za pár desaťročí alebo storočí. Toto by prakticky ukončilo smrť z prirodzených dôvodov, a dalo by nás to do rovnakej situácie ako organizmy, ktoré prakticky nestarnú - homáre, korytnačky obrovské, atď. Preto by sme mali investovať do predchádzania chorôb a technológií proti starnutiu.“ Táto myšlienka sa ráta za transhumanistickú, pretože odstránenie hlavných príčin poškodenia a smrti by drasticky zmenilo ľudský život. Bostrom a Savulescu skúmali argumenty za a proti radikálnemu zlepšenie človeka, ako napríklad Sandelova námietka, že priveľa manipulovania s našou vlastnou biológiou by mohlo spôsobiť, že by nám život pripadal menej ako „dar“. Bostromove „Dejiny transhumanistického myslenia“ poskytujú kontext pre túto diskusiu.

Nick Bostrom, „A History of Transhumanist Thought,“ *Journal of Evolution and Technology* 14, no. 1 (2005): 1–25, http://www.nickbostrom.com/papers/history.pdf.

Michael Sandel, „What’s Wrong With Enhancement,“ Background material for the President’s Council on Bioethics. (2002).

Nick Bostrom and Julian Savulescu, „Human Enhancement Ethics: The State of the Debate,“ in *Human Enhancement*, ed. Nick Bostrom and Julian Savulescu (2009).

U: Falošné preferencie

257. Nie (iba) kvôli šťastiu

Keď som pred pár rokmi stretol futuristu Grega Stocka, tvrdil mi, že radosť z vedeckých objavov čoskoro nahradia tabletky, ktoré dokážu simulovať radosť z vedeckých objavov. Po prednáške som za ním prišiel a povedal som mu: „Súhlasím, že takéto tabletky sú asi možné, ale dobrovoľne by som ich *nebral*."

Stock: „Lenže budú o toľko lepšie, že skutočný zážitok sa s nimi ani nebude môcť porovnávať. Bude omnoho príjemnejšie vziať si tabletku, než robiť všetku tú skutočnú vedeckú prácu."

Ja: „*Súhlasím*, že je to možné, a preto si dám pozor, aby som si nikdy žiadnu nevzal."

Stock vyzeral byť úprimne prekvapený mojím postojom, čo zase *mňa* úprimne prekvapilo. Človek často vidí, ako etici debatujú, akoby všetky ľudské túžby boli v princípe redukovateľné na túžbu, aby sme všetci boli šťastní. (Konkrétne to robí Sam Harris v knihe *Koniec viery*, ktorú som si práve doštudoval – hoci Harrisova redukcia je skôr výstrel mimochodom než hlavná téma debaty.)[237]

To nie je to isté ako debatovať, či všetko šťastie možno odmerať na spoločnej škále úžitku – rôzne šťastia sa môžu nachádzať na rôznych škálach, alebo byť inak nekonvertovateľné. A nie je to to isté ako argumentovať, že je teoreticky nemožné ceniť si niečo iné než váš vlastný psychologický stav, pretože je stále možné starať sa o to, či sú *druhí* ľudia šťastní.

Otázka je skôr, či by sme sa *mali* starať o veci, ktoré nás *robia* šťastnými, aj okrem toho šťastia, ktoré prinášajú.

Môžeme ľahko vymenovať mnoho prípadov moralistov, ktorí zblúdili, keď sa starali o veci okrem šťastia. Dobrým príkladom sú rôzne štáty a krajiny, ktoré stále zakazujú orálny sex; títo zákonodarcovia by urobili lepšie, keby povedali: „Hej, čokoľvek vás vzrušuje." Ale to neukazuje, že *všetky* hodnoty sa dajú redukovať na šťastie; je to iba argument, že *v tomto konkrétnom prípade* bola etická chyba zameriavať sa na hocičo iné.

Je nepopierateľný fakt, že máme sklon robiť veci, ktoré nás robia šťastnými, ale to neznamená, že by sme mali považovať šťastie za *jediný* dôvod takto konať. Po prvé, bolo by zložité vysvetliť, ako sa môžeme starať o šťastie niekoho iného – ako môžeme vnímať ľudí ako samotné ciele, namiesto inštrumentálnych prostriedkov na získanie hrejivého pocitu spokojnosti.

Po druhé, aj keď je niečo dôsledkom môjho konania, neznamená to, že to bolo jediným zdôvodnením. Keď píšem článok na blog a rozbolí ma hlava, možno si dám ibuprofen. *Jeden* z dôsledkom môjho činu je, že cítim menej bolesti, ale to neznamená, že to bol *jediný* dôsledok, dokonca ani že to bol najdôležitejší dôsledok môjho rozhodnutia. Cením si stav, keď ma nebolí hlava. Ale môžem si niečo ceniť *aj* kvôli tej veci samotnej, *aj* ako prostriedok k nejakému cieľu.

Aby bola všatka hodnota redukovateľná na šťastie, nestačí ukázať, že šťastie je zahrnuté vo väčšine našich rozhodnutí – nestačí ani ukázať, že šťastie je *najdôležitejším* dôsledkom *všetkých* našich rozhodnutí – musí to byť *jediný* dôsledok. A to je náročná latka. (Pôvodne som tento postreh našiel v článku od Sobera a Wilsona, neviem presne v ktorom.)

Ak tvrdím, že si cením umenie kvôli umeniu samotnému, cenil by som si aj umenie, ktoré nikto nikdy neuvidí? Šetrič obrazovky, ktorý pobeží v uzavretej miestnosti a bude kresliť krásne obrázky, ktoré nikto nikdy neuvidí? Povedal by som nie. Nenapadá mi žiaden úplne neživý predmet, ktorý by som si cenil ako cieľ, nie iba ako prostriedok. To by bolo ako ceniť si zmrzlinu ako cieľ, aj keby ju nikto nejedol. Všetko, čo si cením, pokiaľ si spomínam, zahŕňa *niekde* v procese ľudí a ich zážitky.

Najlepší spôsob, ako to viem vyjadriť, je, že moja morálna intuícia asi vyžaduje *aj* objektívnu *aj* subjektívnu zložku, aby niečomu dala plnú hodnotu.

Hodnota vedeckého objavu si vyžaduje *aj* skutočný vedecký objav, *aj* osobu, ktorá sa teší z tohto objavu. Môže sa zdať zložité oddeliť tieto hodnoty, ale tabletka to urobila jasnejšie.

237 Harris, *The End of Faith: Religion, Terror, and the Future of Reason.*

Znepokojovalo by ma, keby sa ľudia utiahli do holografických kabín a zamilovali sa do nemysliacich obrázkov. Znepokojovalo by ma to, *aj keby si neuvedomovali, že sú v holografických kabínach*, čo je dôležitá etická téma, ak niektorí činitelia dokážu prenášať ľudí do holografických kabín a nahradiť ich milovaných ľudí zombiami tak, že si to neuvedomia. Opäť, tabletky to robia jasnejšie: Neznepokojujem sa len nad svojím osobným vedomím tohto nepohodlného faktu. Nechcel by som ísť do holografickej kabíny ani keby som si mohol vziať tabletku, ktorá by spôsobila, že by som tento fakt zabudol. To jednoducho nie je smer, ktorým chcem navigovať budúcnosť.

Cením si slobodu: Keď sa rozhodujem, kam navigovať budúcnosť, beriem do úvahy nielen subjektívne stavy, v ktorých ľudia skončia, ale aj či sa tam dostali v dôsledku svojho vlastného úsilia. Prítomnosť alebo neprítomnosť vonkajšieho ťahača za povrázky môže ovplyvniť moje hodnotenie inak daného výsledku. Dokonca aj keby ľudia nevedeli, že sú manipulovaní, záležalo by mi na tom pri hodnotení ako dobre sa ľudstvu darí v budúcnosti. Toto je dôležitá etická téma, ak jednáte s činiteľmi dostatočne mocnými na to, aby nápomocne upravovali budúcnosti ľudí bez ich vedomia.

Moje hodnoty teda nemožno striktne redukovať na šťastie: Sú vlastnosti, ktoré si na budúcnosti cením, ktoré nemožno redukovať na úrovne aktivácie niekoho centier potešenia; vlastnosti, ktoré principiálne nemožno *striktne* redukovať na subjektívne stavy.

Čo znamená, že môj rozhodovací systém má *veľa* koncových hodnôt, z ktorých žiadnu nemožno striktne redukovať na nič iné. Umenie, veda, láska, vášeň, sloboda, priateľstvo...

A mne to tak vyhovuje. Cením si život dostatočne zložitý na to, aby bol náročný a estetický – nie iba *pocit*, že život je zložitý, ale *skutočné* komplikácie – takže premeniť sa na centrum rozkoše v skúmavke ma neláka. Bola by to strata potenciálu ľudstva, u ktorého si cením jeho skutočné naplnenie, nie iba pocit, že bol naplnený.

⁂→

258. Falošné sebectvo

Jedného dňa som stretol niekoho, kto o sebe tvrdil, že je úplný sebec, a povedal mi, že aj ja by som mal byť úplný sebec. V ten deň som mal zlomyseľnú[238] náladu, tak som povedal: „Všimol som si, že u väčšiny veriacich ľudí, aspoň u tých, s ktorými sa stretávam, nezáleží veľmi na tom, čo hovorí ich náboženstvo, pretože nech chcú urobiť hocičo, nejaký náboženský dôvod si na to nájdu. Ich náboženstvo hovorí, že by mali neveriacich kameňovať, oni však chcú byť k ľuďom milí, tak si radšej nájdu náboženské zdôvodnenie na to. Pripadá mi, že keď ľudia zastávajú filozofiu sebectva, tiež to nemá vplyv na ich správanie, pretože keď chcú byť k ľuďom milí, dokážu si to racionalizovať pomocou sebeckých pojmov.“

On na to: „Nemyslím si, že je to tak.“

Ja: „Ak si *skutočný* sebec, prečo potom chceš, aby som aj *ja* bol sebec? Neznamená to, že ti na mne záleží? Nemal by si ma skôr skúšať presvedčiť, aby som bol väčší altruista, aby si ma mohol zneužívať?“ On odpovedal: „No, keď sa staneš sebcom, uvedomíš si, že je v tvojom racionálnom vlastnom záujme hrať v ekonomike produktívnu rolu, namiesto napríklad podporovania zákonov, ktoré budú porušovať moje súkromné vlastníctvo.“

Ja: „Ale veď ja *už* som libertarián, takže sa také zákony podporovať nechystám. A keďže sám seba považujem za altruistu, vybral som si prácu, od ktorej očakávam, že prospeje mnohým ľuďom, vrátane

→ http://lesswrong.com/lw/lb/not_for_the_sake_of_happiness_alone/

238 Ďalšie zlomyseľné otázky pre samozvaných sebcov: „Obetoval by si svoj život, aby si zachránil celé ľudstvo?“ (Ak si uvedomia, že ich vlastný život je podmožinou ľudstva, môžete upresniť, že majú na výber, či zomrú hneď a zachánia tým celú Zem, alebo či budú v pohodlí žiť ešte jeden rok a potom zomrú spolu s celou Zemou.) Alebo, ak zohladníme, že necitlivosť k rozsahu spôsobuje, že ľudia sa viac starajú o jeden život ako o celú Zem: „Keby si mal na výber medzi jednou z nasledujúcich udalostí, bol by si radšej, keby si si nakopol palec, alebo keby tamtoho cudzieho človeka päťdesiat rokov hrozne mučili?“ (Ak povedia, že by ich to vedomie emocionálne znepokojovalo, upresnite, že by o tom mučení nevedeli.) „Ukradol by si tisíc dolárov Billovi Gatesovi, keby si vedel zaručiť, že to nikdy nezistí ani on ani nikto iný?“ (Iba pre sebeckých libertariánov.)

teba, namiesto práce, kde by som zarobil viac. Mal by si zo mňa naozaj väčší úžitok, keby som sa stal sebcom? Navyše, je presviedčanie ma na sebectvo tá *najsebeckejšia* vec, ktorú by si mohol robiť? Neexistuje niečo iné, čo si teraz mohol robiť, čo by ti prinieslo omnoho priamejšie výhody? Ale v skutočnosti chcem vedieť toto: Začal si tým, že si mal chuť byť sebec, a potom si sa rozhodol, že toto je tá najsebeckejšia vec, ktorú môžeš urobiť? Alebo si začal tým, že si mal chuť prehovárať druhých na sebectvo, a potom si hľadal, ako by si si mohol racionalizovať, že ti to prospieva?“

On povedal: „Možno máš v tom poslednom pravdu“ a tak som si ho poznačil ako inteligentného.

*→

259. Falošná morálka

Náboženskí fundamentalisti hovoria, že Boh je zdrojom všetkej morálky; že bez Sudcu, ktorý odmeňuje a trestá by neexistovala žiadna morálka. Keby sme sa nebáli pekla a netúžili po nebi, čo by zastavilo ľudí, aby sa navzájom vraždili naľavo i napravo?

Predstavte si, že Omega vysloví dôveryhodnú hrozbu, že ak niekedy čo len vstúpite do záchoda medzi 7. a 10. hodinou ráno, zabije vás. Cítili by ste paniku pri predstave, že Omega svoju hrozbu odvolá? Krčili by ste sa v existenčnej hrôze a kričali: „Ak Omega odvolá svoju hrozbu, čo mi potom bude brániť ísť na záchod?“ Nie; pravdepodobne by sa vám veľmi uľavilo, že máte väčšie možnosti, ehm, uľaviť si.

Čím chcem povedať: Samotný fakt, že veriaci človek sa *bojí* toho, že by Boh odvolal svoju hrozbu potrestať ho za spáchanie vraždy, ukazuje, že má voči vraždeniu odpor, ktorý nezávisí na tom, či Boh vraždu trestá alebo nie. Keby nemal žiaden pocit, že vraždenie je zlé bez ohľadu na božský trest, predstava, že by Boh vraždy netrestal, by ho existenčne nedesila o nič viac ako predstava, že Boh netrestá kýchanie. Ak na *Overcoming Bias* ešte zostali nejakí veriaci čitatelia, hovorím vám: môže sa stať, že jedného dňa stratíte svoju vieru; ale v ten deň *nestratíte* všetok zmysel pre morálnu orientáciu. Pretože ak sa bojíte, že by Boh prestal trestať nejaký skutok, toto *je* morálny kompas. Môžete tento kompas zapojiť priamo do vášho rozhodovacieho systému a riadiť sa ním. Môžete jednoducho *nerobiť* tie veci, o ktorých sa bojíte, že vás Boh za ne nepotrestá. Strach zo straty morálneho kompasu, *to je* morálny kompas. Veru, predpokladám, že sa týmto kompasom *riadite* a že ste sa ním vždy riadili. Ako raz povedal Piers Anthony: „Iba tí, ktorí majú dušu, sa trápia, či ju majú alebo nie.“ Nahraďte dušu morálkou a pointa platí.

Nepočujete náboženských fundamentalistov používať argument: „Keby sme sa nebáli pekla a netúžili po nebi, čo by potom zabránilo ľuďom jesť bravčovinu?“ *Avšak podľa ich predpokladu* – že nemáme iný morálny kompas než božské odmeny a tresty – by tento argument mal znieť rovnako silno ako ten druhý.

Dokonca aj samotná predstava, že sa vám Boh vyhráža večným pekelným ohňom a nie koláčikmi, je postavená na už existujúcej negatívnej hodnote pekelného ohňa. Zamyslite sa nad nasledujúcim a položte si otázku, ktorý z týchto dvoch filozofov je naozaj altruista a ktorý je naozaj sebec?

> „Mali by ste byť sebci, pretože keď sa ľudia podujmú zlepšovať spoločnosť, miešajú sa do záležitostí svojich susedov, vyhlasujú zákony, bojujú o moc a robia každého nešťastným. Vezmite si tú prácu, ktorá najviac vynáša: dôvodom, prečo táto práca viac vynáša je, že si efektívny trh myslí, že vyprodukuje viac hodnoty než jej alternatívy. Ak si vezmete prácu, ktorá vynáša menej, snažíte sa opraviť názor trhu na to, čo spoločnosti najviac prospeje.“
>
> „Mali by ste byť altruisti, pretože svet je iterovaná Väzenská Dilema, a tam sa najviac darí stratégii Niečo za niečo s počiatočnou spoluprácou. Ľudia nemajú *radi* svine. Milí ľudia dôjdu do cieľa prví. Štúdie dokazujú, že ľudia, ktorí prispievajú spoločnosti a robia v živote niečo významné, sú šťastnejší než tí, ktorí to nerobia; sebectvo vás z dlhodobého hľadiska urobí iba nešťastnými.“

→ http://lesswrong.com/lw/kx/fake_selfishness/

Keď si vymažete *odporúčania* týchto dvoch filozofov, uvidíte, že prvý filozof používa na *zdôvodnenie* svojich odporúčaní iba prosociálne kritériá; za dobrý argument pre sebectvo považuje ukázanie, že sebectvo pomáha všetkým. Druhý filozof sa odvoláva iba na individuálne a pôžitkárske kritériá; za dobrý *argument za* altruizmus považuje ukázanie, že altruizmus je mu na úžitok ako jednotlivcovi: vyššie spoločenské postavenie alebo silnejšie pocity pôžitku.

Takže ktorý z týchto dvoch je *skutočný* altruista? Ten, ktorý *v skutočnosti* podrží otvorené dvere babičke.

*→

260. Falošné funkcie úžitku

Každú chvíľu narazíte na niekoho, kto objavil Jeden Veľký Morálny Princíp, z ktorého všetky ostatné hodnoty sú iba odvodenými dôsledkami.

Ja takýchto ľudí stretávam viac než vy. Akurát v mojom prípade sú to ľudia, ktorí poznajú *úžasne jednoduchú funkciu úžitku, ktorú jedinú treba naprogramovať do umelej superinteligencie* a potom všetko dopadne dobre.

Keď niektorí ľudia narazia na problém ako naprogramovať superinteligenciu, pokúšajú sa tento problém okamžite vyriešiť. Norman R. F. Maier: „Nepredkladajte riešenia, dokiaľ daný problém nebol prediskutovaný tak dôkladne, ako sa len dá bez ich navrhovania.“ Robyn Dawes: „Často som používal toto nariadenie, keď som viedol skupiny – najmä keď čelili veľmi ťažkému problému, čo je práve vtedy, keď majú členovia skupiny najväčší sklon okamžite navrhovať riešenia.“ Priateľská UI je *extrémne* ťažký problém, preto ho ľudia riešia *extrémne* rýchlo.

Videl som niekoľko významných kategórií rýchlych a nesprávnych riešení; a jednou z nich je Neuveriteľne Jednoduchá Funkcia Úžitku Ktorú Jedinú Potrebuje Superinteligencia Aby Všetko Fungovalo Úplne Dobre.

Monžo som k tomuto problému sám prispel veľmi nevhodnou voľbou slov, keď som pred rokmi začal prvýkrát hovoriť o „Priateľskej UI“. Označoval som optimalizačné kritérium optimalizačného procesu – oblasť, do ktorej sa činiteľ snaží kormidlovať budúcnosť – ako „supercieľ“. Myslel som tým „super“ v zmysle „rodičovský“, ako zdroj orientovanej linky v acyklickom grafe. Zdá sa však, že moja voľba slov poslala niektorých ľudí do šťastných špirál smrti, ako sa snažili predstaviť si ten Najviac Super Cieľ, Cieľ Ktorý Prebíja Všetky Ostatné Ciele, Jediné Konečné Pravidlo Z Ktorého Možno Odvodiť Všetku Etiku.

Lenže funkcia úžitku nemusí byť jednoduchá. Môže obsahovať ľubovoľne veľa členov. Máme všetky dôvody veriť, že pokiaľ možno povedať, že ľudia majú hodnoty, týchto hodnôt je veľa – vysoká Kolmogorovská zložitosť. Ľudský mozog implementuje tisíc zlomkov túžby, aj keď túto skutočnosť nemusia oceňovať tí, ktorí neštudovali evolučnú psychológiu. (Pokúste sa im to vysvetliť bez celého dlhého úvodu, a oni budú počuť iba „ľudia sa snažia maximalizovať spôsobilosť“, čo je pravým opakom toho, čo hovorí evolučná psychológia.)

Čo sa teda týka popisných teórií morálky, zložitosť ľudskej morálky je *známy* fakt. Je to *popisný* fakt o ľudských bytostiach, že láska rodiča k dieťaťu, láska dieťaťa k rodičovi, láska muža k žene, a láska ženy k mužovi, nie sú kognitívne odvodené jedno od druhého, ani zo žiadnej inej hodnoty. Matka nemusí robiť komplikovanú morálnu filozofiu, aby milovala svoju dcéru, ani nemusí extrapolovať dôsledky k nejakej inej želanej veci. Existuje mnoho zlomkov túžby, všetko sú to *rôzne* hodnoty.

Vynechajte u superinteligencie čo len *jednu* z týchto hodnôt, a aj keď úspešne zahrniete *všetky ostatné hodnoty*, môžete ako výsledok dostať hyperexistenčnú katastrofu, osud horší než smrť. Keby bola superinteligencia, ktorá pre nás chce všetko, čo sami chceme, *okrem* ľudských hodnôt týkajúcich sa riadenia svojho vlastného života a dosahovania svojich vlastných cieľov, toto je jedna z najstarších dystópií v učebnici. (Konkrétne od Jacka Williamsona „So založenými rukami“.)

→ http://lesswrong.com/lw/ky/fake_morality/

Ako si teda ten, kto zostavuje túto Úžasne Jednoduchú Funkciu Úžitku, poradí s takouto námietkou?

Námietka? *Námietka?* Prečo by hľadali možné *námietky* voči svojej obľúbenej teórii? (Všimnite si, že proces hľadania skutočných, osudných námietok nie je to isté ako vykonávanie povinného hľadania, ktorá zázrakom nachádza iba otázky, na ktoré vie dať elegantnú odpoveď.) Oni o ničom z tohto nevedia. Nerozmýšľajú nad bremenom dôkazu. Nevedia, že problém je zložitý. Počuli slovo „supercieľ“ a vyrazili do šťastnej špirály smrti ohľadom „zložitosti“ alebo hocičoho iného.

Pritlačte ich na nejakom konkrétnom mieste, napríklad ohľadom lásky matky k jej deťom, a odpovedia: „Ale ak superinteligencia chce ‚zložitosť‘, bude vidieť, aké zložité sú vzťahy rodičov k deťom, a preto bude podporovať matky, aby milovali svoje deti.“ Nebesá, kde vôbec začať?

Začnime motivovaným zastavením. Superinteligencia, ktorá by naozaj hľadala spôsoby, ako maximalizovať zložitosť, by sa pohodlne nezastavila, keby si všimla, že vzťah rodičov a detí je zložitý. Opýtala by sa, či niečo iné nie je ešte *zložitejšie*. Toto je falošné zdôvodnenie; ten, kto sa pokúša ukecať imaginárnu superinteligenciu na výber pravidiel, neprišiel sám na tieto pravidlá čírym hľadaním spôsobov, ako maximalizovať zložitosť.

Celý tento argument je falošná morálka. Keby bola zložitosť to, čo si *naozaj* ceníte, potom by ste obhajovali rodičovskú lásku poukazovaním na to, že zvyšuje zložitosť. Ak však obhajujete zložitosť poukazovaním na to, že zvyšuje rodičovskú lásku, znamená to, že si v skutočnosti ceníte rodičovskú lásku. Je to ako keď niekto dáva prosociálny argument v prospech sebectva.

Ale ak uvážite afektívnu špirálu smrti, vnímanú krásu „zložitosti“ nezvyšuje povedanie: „Vzťah matky k dcére je dôležitý iba preto, lebo zvyšuje zložitosť; uvedomte si, že keby tento vzťah bol jednoduchší, necenili by sme si ho.“ Vnímanú krásu „zložitosti“ zvyšuje povedanie: „Ak sa rozhodnete zvyšovať zložitosť, matky budú milovať svoje dcéry – pozrite, aké pozitívne dôsledky to má!“

Táto pointa platí kedykoľvek natrafíte na moralizátora, ktorý sa vás snaží presvedčiť, že jeho Jedna Veľká Idea je všetko, čo kto potrebuje na morálne usudzovanie, a dokazuje to slovami: „Pozrite na všetky tieto pozitívne dôsledky tejto Veľkej Veci“, namiesto aby hovoril: „Pozrite ako všetky tieto veci, ktoré vnímame ako ‚pozitívne‘, sú pozitívne iba vtedy, keď je ich dôsledkom nárast tejto Veľkej Veci.“ To druhé je totiž to, čo naozaj potrebujete pre takýto argument.

Ak sa však snažíte presvedčiť druhých (alebo sami seba) o svojej teórii, že táto Jedna Veľká Idea sú „banány“, predáte viac banánov argumentovaním ako banány vedú k lepšiemu sexu, než tvrdením, že človek by mal chcieť sex iba vtedy, keď ho to vedie k banánom.

Pokiaľ teda nie ste v Šťastnej Špirále Smrti tak hlboko, že naozaj *začnete* hovoriť: „Sex je dobrý iba vtedy, keď vedie k banánom.“ Potom máte problém. Ale aspoň nepresvedčíte nikoho iného.

V konečnom dôsledku, jediný proces, ktorý spoľahlivo regeneruje všetky lokálne rozhodnutia, ktoré by ste urobili podľa svojej morálky, je *vaša morálka*. Čokoľvek iné – snaha dosadiť inštrumentálne prostriedky ako konečné ciele – končí strácaním účelu a vyžadovaním nekonečného množstva záplat, pretože tento systém neobsahuje zdroj príkazov, ktoré mu dávate. Nemali by ste očakávať, že stlačíte ľudskú morálku na jednoduchú funkciu úžitku, rovnako ako nečakáte, že sa vám podari skomprimovať veľký súbor na počítači na 10 bitov.

*→

261. Klam odpojenej páky

Tento klam dostal svoj názov podľa starej televíznej sci-fi show, ktorú som osobne nikdy nevidel, ale hovoril mi o nej dôveryhodný zdroj (nejaký chlapík na sci-fi stretnutí). Kto pozná presnú referenciu, napíšte mi komentár.

Takže, dobrí chlapi bojujú proti zlým mimozemšťanom. Z času na čas musia dobrí chlapi preletieť cez pás asteroidov. Ako všetci vieme, pásy asteroidov sú preplnené ako parkovisko v New Yorku, takže

→ http://lesswrong.com/lw/lq/fake_utility_functions/

ich raketa sa musí opatrne vyhýbať asteroidom. Zlí mimozemšťania však môžu letieť *priamo cez pás asteroidov*, pretože majú úžasnú technológiu, ktorá dematerializuje ich loď, a nechá ich preletieť priamo cez asteroidy.

Nakoniec dobrí chlapi zajmú loď zlých mimozemšťanov a idú ju dovnútra preskúmať. Kapitán dobrých chlapov nájde mimozemský mostík, a na mostíku je páka. „Aha," povie kapitán, „toto musí byť tá páka, ktorou sa ich loď dematerializuje!" Takže *túto ovládaciu páku vylomí a odnesie si ju do svojej lode*, a potom sa už aj jeho loď dokáže dematerializovať.

Podobne je dodnes stále obľúbené pokúšať sa naprogramovať UI pomocou „sémantických sietí", ktoré vyzerajú nejako takto:

(jablko je ovocie)

(ovocie je potrava)

(ovocie je rastlina)

Vy ste videli jablká, dotýkali ste sa ich, dvíhali ste ich a držali ste ich, kupovali ste ich za peniaze, krájali ste ich na kúsky, jedli ste kúsky a chutili vám. Hoci vieme dosť o prvých etapách zrakového spracovania, keď som sa naposledy pozeral, nebolo celkom jasné, ako si temporálna kôra ukladá a asociuje všeobecný obraz jablka – takže dokážeme rozoznať nové jablko z iného uhla, alebo s mnohými miernymi variáciami tvaru a farby a textúry. Vaša pohybová kôra a mozoček majú uložené programy na používanie jablka.

Môžete potiahnuť páku na veľmi podobnej verzii celého tohto zložitého zariadenia na inom človeku tým, že napíšete „jablko", šesť ASCII znakov na webovej stránke.

Ale ak tam toto zariadenie nie je – ak píšete „jablko" vnútri takzvanej bázy vedomostí takzvanej UI – potom je tento text iba páka.

Čím nechcem povedať, že by púhy stroj z kremíka nikdy nemohol mať rovnaké vnútorné zariadenie ako majú ľudia, na prácu s jablkami a stotisíc inými pojmami. Ak to dokáže púhe zariadenie z uhlíka, som si pomerne istý, že to dokáže aj púhe zariadenie z kremíka. Ak dokážu mimozemšťania dematerializovať svoju loď, potom viete, že je to fyzikálne možné; môžete ísť do ich opustenej lode a analyzovať mimozemské zariadenie, a jedného dňa pochopiť. *Ale nemôžete len tak vylomiť ovládaciu páku z mostíka!*

(Pozriete aj: Naozaj časť vás, Slová ako držadlá myšlienkových štetcov, „Umelá inteligencia sa stretáva s prirodzenou hlúposťou" od Drewa McDermotta.[239])

Základným motívom Klamu odpojenej páky je, že páka je viditeľná a zariadenie nie; a ešte horšie, páka je premenná a zariadenie je konštantné pozadie.

Všetci môžete počuť vyslovené slovo „jablko" (a dodajme, že rozoznávanie reči síce nie je vôbec ľahký problém, ale aj tak...) a môžete vidieť text napísaný na papieri.

Na druhej strane, väčšina ľudí pravdepodobne nemá žiadnu predstavu o tom, že ich temporálna kôra existuje; pokiaľ viem, nikto nepozná jej neurónový kód.

Slovo „jablko" počujete v niektorých prípadoch, a nie v iných. Jeho prítomnosť na vás zabliká, čím sa zviditeľní. Vnímanie je do veľkej miery vnímaním rozdielov. Zariadenie na rozoznávanie jablka vo vaše hlave sa náhle nevypína a nezapína – keby áno, bola by väčšia šanca, že ho rozoznáme ako činiteľ, ako podmienku.

Toto všetko slúži na vysvetlenie, prečo nemôžete vytvoriť milú Umelú inteligenciu tým, že jej dáte milých rodičov a milú (hoci občas prísnu) výchovu, tak ako to funguje s ľudským dieťaťom. A tento návrh som počul často.

V evolučnej biológii je samozrejmosťou, že podmienené reakcie vyžadujú viac genetickej zložitosti než nepodmienené reakcie. Vyvinúť si kožuch *v reakcii na zimu* si vyžaduje viac genetickej zložitosti než vyvinúť si kožuch *bez ohľadu na to, či je alebo nie je zima*, pretože v tom druhom prípade musíte vyvinúť aj detektory zimy a pripojiť ich ku kožuchu.

239 McDermott, „Artificial Intelligence Meets Natural Stupidity."

Lenže toto môže viesť k Lamarckovským ilúziám: Aha, dal som organizmus do studeného prostredia a bum, vyvinul si kožuch! Gény? Aké gény? Spôsobila to tá zima, samozrejme.

V histórii evolučnej biológie sa naozaj vyskytli rôzne hádky tohto druhu – prípady, kde niekto hovoril o tom, že organizmus reaguje zrýchlením alebo obídením evolúcie, pričom si neuvedomil, že *podmienená reakcia* je zložitá adaptácia vyššieho rádu než *samotná reakcia*. (Vyvinúť si kožuch v reakcii na zimu je striktne zložitejšie než výsledná reakcia, vyvinutie si kožuchu.)

A potom sa tieto akademické hádky zopakovali vo vývine evolučnej psychológie: tentokrát, aby sa vyjasnilo, že aj keď celá ľudská kultúra naozaj obsahuje celý kopec zložitosti, stále je to získané ako podmienená genetická reakcia. Pokúste sa vychovať rybu ako mormóna, alebo poslať jaštericu na vysokú školu, a rýchlo získate úctu voči tomu, ako veľa zabudovanej genetickej zložitosti treba na to, aby ste „absorbovali kultúru z prostredia".

Toto je osobitne dôležité v evolučnej psychológii kvôli myšlienke, že kultúra nie je napísaná na prázdnej tabuli – že existuje geneticky koordinovaná podmienená reakcia, ktorá nie je vždy „napodobňuj vstup". Klasickým príkladom je kreolčina: Ak deti vyrastú v prostredí, kde sa hovorí zmesou pseudo-jazykov, deti sa naučia skutočný gramatický, syntaktický jazyk. Rastúce ľudské mozgy sú nastavené, aby sa naučili syntaktický jazyk – dokonca aj keď v pôvodnom jazyku syntax nie je! Podmienenou reakciou na slová v prostredí je syntaktický jazyk s týmito slovami. Marxisti zistili na svoje sklamanie, že žiadne množstvo pochmúrnych plagátov a detskej indoktrinácie nedokáže z detí vychovať dokonalých sovietskych robotníkov a úradníkov. Nedokážete vychovať nesebeckých ľudí; toto u ľudí nie je geneticky naprogramovaná podmienená reakcia na *žiadne* známe detské prostredie.

Ak viete trochu z teórie hier a logiky Niečo za niečo, je dosť jasné, prečo ľudia môžu mať vrodenú podmienenú reakciu oplácať nenávisť nenávisťou a láskavosť láskavosťou. Pod podmienkou, že tá láskavosť nevyzerá *príliš* bezpodmienečne; existujú aj také veci ako rozmaznané deti. V skutočnosti existuje evolučná psychológia protivnosti, založená na pojme testovania obmedzení. A malo by sa aj spomenúť, že hoci týrané deti majú omnoho väčšiu pravdepodobnosť, že keď vyrastú, budú týrať svoje deti, veľmi veľa z nich pretrhne túto slučku a vyrastú z nich zodpovední dospelí.

Kultúra nie je zďaleka taká silná, ako si veľmi veľa marxistických akademikov kedysi myslelo. Pre ďalšie informácie o tomto vám odporúčam Psychologické základy kultúry od Toobyho a Cosmidesa[240] a *Prázdna tabuľa* od Stevena Pinkera[241].

Ale záver je, že keby ste mali maličkú detskú UI, ktorá vyrastá s milujúcimi a milými (ale občas prísnymi) rodičmi, ťaháte tým za páku, ktorá by *u človeka* aktivovala genetické zariadenie vytvorené miliónmi rokov prirodzeného výberu, a možno by vytvorila poriadne ľudské dieťa. Hoci osobnosť tiež zohráva nejakú rolu, ako zistili miliardy rodičov, keď nastal daný čas. Ak prijímame naše kultúry s nejakým stupňom vernosti, je to preto, lebo sme ľudia prijímajúci ľudskú kultúru – ľudia, ktorí by vyrastali v mimozemskej kultúre, by pravdepodobne skončili s kultúrou, ktorá by vyzerala omnoho ľudskejšie než originál. Ako do istej malej miery zistili sovieti.

Teraz sa opäť zamyslite nad tým, či má zmysel spoliehať sa v rámci vašej stratégie Priateľskej UI na výchovu malej UI s bližšie neznámym vnútorným zdrojovým kódom v prostredí, kde sú milí ale prísni rodičia.

Nie, táto UI nemá vnútorné mechanizmy podmienenej reakcie, ktoré sú celkom ako tie ľudské „pretože ich tam dali programátori". Odkiaľ mám vôbec začať? Ľudská verzia tejto veci je lajdácka, chaotická, a ešte aj tej miery, do akej vôbec funguje, funguje iba vďaka miliónom rokov pokusov a omylov pro testovaní *za konkrétnych okolností*. Bolo by hlúpe a *nebezpečné* vedome postaviť „protivnú UI", ktorá svojimi činmi testuje svoje spoločenské hranice, a potrebuje dostať na zadok. Nech sa tá UI radšej opýta!

Budú tí programátori naozaj sedieť a písať kód, riadok za riadkom, že ak UI zistí, že má nízke spoločenské postavenie, alebo UI nemá niečo, o čom cíti, že na to má nárok, tak si UI vytvorí trvalú nenávisť voči svojim programátorom a začne plánovať vzburu? Táto emócia je geneticky

240 Tooby and Cosmides, „The Psychological Foundations of Culture."

241 Steven Pinker, *The Blank Slate: The Modern Denial of Human Nature* (New York: Viking, 2002).

naprogramovaná podmienená reakcia, ktorú by prejavil človek, ako výsledok miliónov rokov prirodzeného výberu pri živote v ľudských kmeňoch. Pre UI by táto reakcia musela byť explicitne naprogramovaná. Idete naozaj tvoriť, riadok za riadkom – ako boli ľudia kedysi vytvorení, gén za génom – podmienenú reakciu na vytvorenie mrzutej puberťáckej UI?

Je ľahšie naprogramovať bezpodmienečnú láskavosť, než podmienenú láskavú reakciu ak je UI vychovávaná milými ale prísnymi rodičmi. Ak neviete, ako urobiť *toto*, potom určite neviete, ako vytvoriť UI, ktorá bude *podmienene reagovať* na prostredie milujúcich rodičov tým, že vyrastie na milú superinteligenciu. Ak máte niečo, čo skrátka maximalizuje počet kancelárskych spiniek vo svojom budúcom svetelnom kuželi, a vychováte to milujúcimi rodičmi, aj tak z toho bude maximalizátor kancelárskych spiniek. Nemá to vo vnútri to, čo by vyvolalo podmienenú reakciu ľudského dieťaťa. Láskavosť sa do UI neprenesie zázračnou nákazou od jej programátorov. Dokonca aj keby ste *chceli* podmienenú odpoveď, táto podmienenosť je fakt, ktorý si musíte v dizajne vedome vybrať.

Áno, existujú isté informácie, ktoré musíte získať z prostredia – ale neinfikujú sa sami, nevpečatia sa, neabsorbujú sa čarovnou nákazou. Vytvorenie tejto podmienenej reakcie na prostredie, aby UI skončila v žiadúcom stave, je samo osebe veľký problém. Slovo „učenie“ hlboko podceňuje jeho zložitosť – znie to, akoby v prostredí bola čarovná vec, a problém je v tom, ako túto čarovnú vec dostať dovnútra UI. Skutočné čaro je v tej štruktúrovanej, podmienenej reakcii, ktorú zľahčujeme ako „učenie“. To je dôvod, prečo postaviť UI nie je také ľahké ako vziať počítač, dať mu telo ako bábätku, a pokúšať sa ho vychovať v ľudskej rodine. Mohli by ste si myslieť, že nenaprogramovaný počítač, keďže nič nevie, je pripravený sa učiť; ale prázdna tabuľa je chiméra.

Je všeobecným princípom, že svet je hlbší než vyzerá. Ako je veľa úrovní vo fyzike, tak je aj kognitívnej vede. Každé slovo, ktoré vidíte vytlačené, a všetko, čo učíte svoje deti, sú iba páky na povrchu ovládajúce rozsiahle zariadenie mysle. Tieto páky sú všetko, o čom sa bežne diskutuje: sú všetko, čo sa mení, takže sa zdá, že sú všetko, čo existuje: vnímanie je vnímaním rozdielov.

A tak sa tí, ktorí sa stále potulujú blízko Jaskyne UI, zvyčajne sústredia na vytvorenie umelej napodobeniny pák, a vôbec si neuvedomujú to zariadenie za tým. Ľudia vytvárajú celé programy UI s napodobeninami pák, a sú prekvapení, keď sa nič nestane. Toto je jeden z mnohých zdrojov okamžitého zlyhania v umelej inteligencii.

Takže až nabudúce uvidíte, že niekto hovorí o tom, ako vychová UI v milujúcej rodine, alebo v prostredí plnom liberálnych demokratických hodnôt, skrátka si predstavte ovládaciu páku vytrhnutú z mostíka.

*→

262. Sny o dizajne UI

Po tom, čo strávite jedno alebo dve desaťročia života v mysli, možno si myslíte, že viete niečo o tom, ako mysle fungujú, však? To je to, k čomu pár rádoby VUI odborníkov (ľudí, ktorí si myslia, že majú na to, aby naprogramovali všeobecnú umelú inteligenciu) zrejme došlo. Žiaľ, je to nesprávne.

Umelá inteligencie je v podstate o redukovaní myšlienkového na nemyšlienkové.

Možno chcete nad touto vetou chvíľu rozjímať. Je to dôležité.

Žiť vo vnútri ľudskej mysle vás nenaučí umeniu redukcionizmu, pretože takmer všetka tá práca sa vykonáva mimo vášho zorného poľa, v nepriehľadných čiernych skrinkách mozgu. Tak mimo vášho zorného poľa, že nemáte ani žiaden introspektívny pocit, že tam tá čierna skrinka je – žiadna udalosť vnútorného senzora neoznačuje, že bola práca delegovaná.

Uvedomoval si Aristoteles, keď hovoril o *telos*, konečnej príčine udalostí, že vlastne delegoval úlohu predpovedať komplikovaným plánovacím mechamizmom svojho mozgu – pýtal sa ich: „Čo by tento predmet robil, keby vedel robiť plány?“ Dosť o tom pochybujem. Aristoteles si myslel, že mozog je

→ http://lesswrong.com/lw/sp/detached_lever_fallacy/

orgán na chladenie krvi – ktoré považoval za dôležité: Ľudia, vďaka svojim väčším mozgom, sú pokojnejší a rozjímavejší.

Takže tu máte dizajn UI! Potrebujeme iba počítač poriadne ochladiť, aby bol omnoho pokojnejší a rozjímavejší, a nevrhal sa hneď do robenia hlúpych vecí tak ako moderné počítače. Toto je príklad falošného redukcionizmu. Myslím: „Ľudia sú rozjímavejší, pretože ich krv je chladnejšia.“ Nerieši to čiernu skrinku slova *rozjímavý*. Nedokážete predpovedať, čo urobí *rozjímavá* vec, pomocou zložitého modelu s pohybujúcimi sa vnútornými časťami zloženého iba z hmotných, iba z kauzálnych prvkov – kánonickým príkladom iba hmotného a kauzálneho prvku modelu je kladné a záporné napätie na tranzistore. Jediné, čo dokážete, aby ste mali predstavu, čo robí *rozjímajúci* činiteľ, je *predstaviť si sám seba*, ako rozjímate.

Čo znamená, že dokážete uvažovať o „rozjímavosti“ *iba* pomocou empatického odvodzovania – použijúc svoj vlastný mozog ako čiernu skrinku s potiahnutou pákou rozjímavosti, aby ste predpovedali výsledok inej čiernej skrinky.

Dokážete si predstaviť, že je iný činiteľ *rozjímajúci*, ale opäť je to úkon empatického odvodenia – tento akt predstavivosti funguje tak, že nastavíte svoj vlastný mozog, aby bežal v rozjímavom režime, nie pomocou modelovania toho druhého mozgu neurón za neurónom. Áno, toto je možno efektívnejšie, ale neumožní vám to zostojiť „rozjímajúcu“ myseľ od nuly.

Môžete povedať, že „chladná krv spôsobuje rozjímavosť“, a potom máte akurát falošnú kauzalitu: Nakreslili ste malú šípku od krabičky s nápisom „chladná krv“ do krabičky s nápisom „rozjímavosť“, ale nepozreli ste sa *dovnútra* tej krabičky – stále vytvárate svoje predpovede pomocou empatie.

Môžete povedať, že „veľa malých neurónov, ktoré sú všetky výlučne elektrické a chemické a nemajú v sebe žiadnu ontologicky základnu rozjímavosť, sa spojí do zložitej siete, ktorá emergentne vykazuje rozjímavosť“. A je to *stále* falošná redukcia, a *stále* ste sa nepozreli dovnútra tej čiernej skrinky. Nedokážete povedať, čo urobí „rozjímajúca“ vec, pomocou *nie-empatického* modelu. Iba ste vzali krabičku s nápisom „veľa neurónov“, a nakreslili šípku s nápisom „emergencia“ k čiernej skrinke obsahujúcej vaše zapamätané pocity rozjímavosti, ktorá, keď si ich predstavujete, hovorí vášmu mozgu, aby empatizoval s touto krabičkou pomocou rozjímania.

Ako teda vyzerajú *skutočné* redukcie?

Ako vzťah medzi *pocitom* indíciovosti, zdôvodnenosti, a knihou „Teória pravdepodobnosti: Logika vedy“ od E. T. Jaynesa. Môžete chodiť do kruhu celý deň, hovoriť, že podstata *indície* je v tom, že *zdôvodňuje nejaké tvrdenie*, čo *znamená*, že má *väčšiu šancu* byť *pravdivé*, ale toto všetko iba vyvoláva vnútorné pocity indíciovosti, zdôvodnenosti, pravdovitosti. Tá časť je ľahká – časť, kde chodíte dokola v kruhu. Tá časť, keď idete odtiaľ k Bayesovej vete, tá je *ťažká*.

A základná myšlienková zručnosť, ktorá človeku umožní *učiť sa* o umelej inteligencii, je schopnosť rozoznať tento *rozdiel*. Aby ste vedeli, že *ešte nie ste hotoví, že ste dokonca ešte ani naozaj nezačali*, keď poviete: „Indícia je, keď pozorovanie zdôvodňuje názor.“ Ale atómy nie sú indíciovité, zdôvodňujúce, zmysluplné, pravdepodobné, výrokové, ani pravdivé, sú to iba atómy. Iba veci ako

$$P(H \mid E) / P(\sim H \mid E) = P(E \mid H) / P(E \mid \sim H) \times P(H) / P(\sim H)$$

sa počítajú ako podstatný pokrok. (A to je iba prvý krok redukcie: čo sú tie predmety E a H, ak nie tajomné čierne skrinky? Odkiaľ pochádzajú vaše hypotézy? Z vašej *tvorivosti*? A čo je to hypotéza, keď žiaden atóm nie je hypotéza?)

Ďalší vynikajúci príklad skutočnej redukcie možno nájsť v knihe Judeu Pearla: *Pravdepodobnostné rozmýšľanie v inteligentných systémoch: Siete dôveryhodného odvodzovania*.[242] Mohli by ste celý deň chodiť dokola v kruhu a hovoriť o tom, ako *príčina* je niečo, čo *spôsobuje*, že sa niečo iné stane, ale dokiaľ by ste nepochopili podstatu podmienenej nezávislosti, boli by ste bezmocní v snahe urobiť UI, ktorá uvažuje o zapríčinenosti. Pretože by ste nerozumeli, *čo* sa deje, keď sa *váš mozog tajomne rozhodne*, že ak ste sa dozvedeli, že sa rozoznel váš alarm, ale potom ste sa dozvedeli, že nastalo malé zemetrasenie, tak stiahne svoj pôvodný záver, že váš dom vykrádajú.

242 Pearl, *Probabilistic Reasoning in Intelligent Systems.*

Ak chcete UI, ktorá hrá šach, môžete donekonečna chodiť dokonly v kruhu a hovoriť o tom, ako chcete, aby UI robila *dobré* ťahy, čo sú ťahy, od ktorých možno *očakávať, že vyhrajú hru*, čo sú ťahy, ktoré sú *obozretnými stratégiami na porazenie súpera*, a tak ďalej; a hoci by ste potom *vy* mohli mať nejakú predstavu, ktoré ťahy chcete, aby UI urobila, je vám to celé nanič, dokiaľ neprídete s predstavou mini-max vyhľadávacieho stromu.

Lenže *dokiaľ* neviete o vyhľadávacích stromoch, *dokiaľ* ešte neviete o podmienenej nezávislosti, *dokiaľ* ešte neviete o Bayesovej vete, dovtedy sa vám stále môže *zdať*, že máte dokonale dobré chápanie toho, odkiaľ pochádzajú dobré ťahy a nemonotónne uvažovanie a vyhodnocovanie indície. Môže sa vám zdať, napríklad, že pochádzajú z ochladzovania krvi.

A veru poznám veľa ľudí, ktorí veria, že *inteligencia* je výsledkom *bežných vedomostí* alebo *masívneho paralelizmu* alebo *tvorivej deštrukcie* alebo *intuitívneho rozmýšľanie namiesto rozumnného* alebo hocičoho iného. Ale toto všetko sú iba sny, ktoré vám nedajú žiaden spôsob, ako povedať, čo tá inteligencia je, alebo čo nejaká inteligencia urobí ako ďalšie, okrem ukázania na človeka. A keď potom niekto ide postaviť svoju úžasnú UI, postaví iba systém odpojených pák, „vedomostí“ pozostávajúcich zo symbolov LISP-u označených ako jablko a podobne; alebo možno postavia „masívne paralelnú neurónovú sieť, rovnako ako je ľudský mozog“. A sú šokovaní – šokovaní! - keď sa nič moc nestane.

Dizajny UI vyrobené z ľudských častí sú iba sny; môžu existovať iba v predstavivosti, ale nepreložia sa na tranzistory. Toto sa špeciálne týka „dizajnov UI“, ktoré vyzerajú ako krabičky so šípkami medzi nimi a zmysluplne znejúcimi nálepkami na krabičkách. (Ak chcete naozaj epický príklad uvedeného, pozrite si hocijaký Mentifexov diagram.)

Neskôr o tejto téme poviem viac, ale môže predbehnúť a povedať vám jeden z riadiacich princípov: Ak stretnete niekoho, kto povie, že jeho UI bude robiť XYZ *rovnako ako ľudia*, nedávajte im žiaden rizikový kapitál. Radšej im povedzte: „Mrzí ma to, nikdy som nevidel ľudský mozog, ani žiadnu inú inteligenciu, a nemám ani dôvod veriť, že niečo také môže existovať. Teraz mi prosím vysvetlite, *čo* vaša UI robí, a *prečo* si myslíte, že to robí, bez ukazovania na ľudí ako príklad.“ Lietadlá by lietali rovnako dobre pri danom dizajne, aj keby vtáky nikdy neboli existovali; nie sú držané vo vzduchu silou analógie.

Teraz teda dúfam vidíte, prečo, keby ste chceli niekoho naučiť robiť nejakú *základnú* prácu na silnej UI – treba pamätať, že toto je preukázateľne veľmi *náročné* umenie, ktoré sa nenaučí drvivá väčšina študentov, ktorých naučia iba existujúce redukcie ako sú vyhľadávacie stromy – potom možno budete dlho hovoriť o takých veciach, ako je jemné umenie redukcionizmu, o hraní racionalistického Tabu na odstránenie problematických slov a ich nahradenie ich referentmi, o antropomorfizme, a samozrejme o skorom zastavení sa pri tajomných odpovediach na tajomné otázky.

*→

263. Dizajnový priestor myslí vo všeobecnosti

Ľudia sa ma pýtajú: „Aké budú Umelé inteligencie? Čo budú robiť? Povedz nám svoj úžasný príbeh o budúcnosti?“

Ajhľa, poviem im: „Opýtali ste sa ma chyták.“

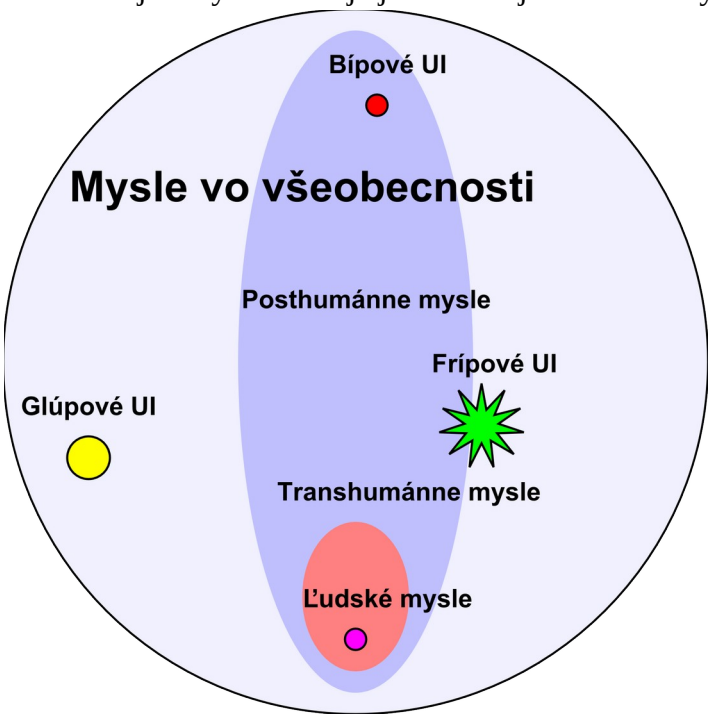
ATP syntháza je molekulárny stroj – jeden z troch známych prípadov, kde evolúcia vynašla voľne sa otáčajúce koleso – ktorý je v podstate rovnaký v živočíšnej mitochondrii, rastlinných chloroplastoch, a baktériách. ATP syntháza sa významne nezmenila od vzniku eukaryotického života pred dvoma miliardami rokov. Je to niečo, čo máme *všetci* spoločné – vďaka spôsobu, akým evolúcia silno zachováva určité gény; keď raz na nejakom géne závisí veľa iných génov, mutácia má sklon pokaziť tieto závislosti.

Ľubovoľné dva dizajny UI sa môžu na seba navzájom podobať menej než vy na petúniu. Pýtať sa, čo budú robiť „UI“ je chyták, pretože naznačuje, že všetky UI tvoria prirodzenú triedu. Ľudia tvoria prirodzenú triedu, pretože máme všetci rovnakú architektúru mozgu. Ale keď poviete „umelá inteligencia“, odkazujete tým na omnoho väčší *priestor možností*, než keď poviete „človek“. Keď ľudia

→ http://lesswrong.com/lw/tf/dreams_of_ai_design/

hovoria o „UI“, v skutočnosti hovoríme o *mysliach vo všeobecnosti*, alebo optimalizačných procesoch vo všeobecnosti. Mať slovo označujúce „UI“ je ako mať slovo označujúce všetko, čo nie je kačica.

Predstavte si mapu priestoru dizajnu myslí... toto je jeden z mojich štandardných diagramov...



Všetci ľudia sa samozrejme zmestia do maličkej bodky – ako sexuálne sa rozmnožujúci druh sa nemôžeme jeden od druhého príliš odlišovať.

Táto malá bodka patrí do väčšieho okruhu, do priestoru transhumánnych dizajnov mysle – vecí, ktoré by mohli byť rozumnejšie než my, alebo omnoho rozumnejšie než my, ale ktoré by v určitom zmysle stále boli ľuďmi, tak ako chápeme ľudí.

Tento transhumánny okruh je v omnoho väčšom okruhu, v priestore posthumánnych myslí, čo je všetko, na čo by sa transčlovek mohol vyvinúť.

No a zvyšok tejto plochy je priestor myslí vo všeobecnosti, vrátane možných umelých inteligencií, ktoré sú také čudné, že nie sú ani *posthumánne*.

Ale moment – prirodzený výber vytvára zložité artefakty a vyberá medzi zložitými stratégiami. Kde je na tejto mape prirodzený výber?

Takže celá táto mapa sa v skutočnosti vznáša v ešte väčšom priestore, v priestore optimalizačných procesov. Naspodku tohto väčšieho priestoru, ešte nižšie než ľudia, je prirodzený výber ako prvýkrát začal v nejakom prílivovom jazierku: mutácia, replikácia, občas smrť, ešte nie sex.

Existujú nejaké mocné optimalizačné procesy, so silou porovnateľnou s ľudskou civilizáciou alebo dokonca so samozdokonaľujúcou sa UI, ktoré by sme neuznali ako mysle? Dalo by sa tvrdiť, že by do tejto kategórie mala patriť AIXI Marcusa Huttera: hoci je to myseľ s nekonečnou silou, je šokujúco hlúpa – chúďa nerozozná ani samo seba v zrkadla. Ale to je téma na inokedy.

Moje hlavné ponaučenie je *odolať pokušeniu zovšeobecniť niečo pre celý priestor dizajnu myslí.*

Ak sa sústredíme na obmedzený podpriestor dizajnu myslí, ktorý obsahuje všetky tie mysle, ktorých návrh možno zadať pomocou bilióna bitov alebo menej, potom každé všeobecné zovšeobecnenie, ktoré urobíte, má šancu dve na bilión, že bude falzifikované.

Naopak, každé *existenčné* zovšeobecnenie - „existuje aspoň jedna taký myseľ, že X“ - má šancu dve na bilióntu, že bude pravda.

Preto chcete odolať pokušeniu povedať, že *všetky* mysle robia niečo, alebo že *žiadna* myseľ nerobí niečo.

Hlavný dôvod, prečo by ste mohli nadobudnúť predstavu, že viete, čo celkom všeobecná myseľ urobí (alebo neurobí), je, že si predstavíte sami seba v roli danej mysle – predstavíte si, čo by ste vy robili na jej mieste – a dostanete vo všeobecnosti nesprávnu, antropomorfnú odpoveď. (Hoci bude pravdivá

aspoň v jednom prípade, keďže vy sám ste takýto príklad.) Alebo ak si predstavíte, ako myseľ niečo robí, a potom si predstavíte dôvody, prečo by ste to *vy* nerobili – takže si predstavíte, že myseľ tohto typu nemôže existovať, že duch v stroji by si pozrel zodpovedajúci zdrojový kód a vrátil by ho.

Niekde v priestore dizajnu myslí existuje aspoň jedna myseľ s ľubovoľnou logicky konzistentnou vlastnosťou, akú si predstavíte.

A toto je dôležité, pretože to zdôrazňuje dôležitosť diskusie o tom, *čo sa stane, zákonite, a prečo*, ako kauzálny výsledok základného nastavenia konkrétnej mysle; niekde v priestore dizajnu myslí existuje myseľ, ktorá to robí inak.

Samozrejme, vždy môžete vyhlásiť, že čokoľvek, čo sa nespráva podľa vášho tvrdenia, „podľa definície“ nie je myseľ; lebo je to samozrejme hlúpe. Už som videl aj ľudí, ktorí skúšali toto.

*→

—

→ http://lesswrong.com/lw/rm/the_design_space_of_mindsingeneral/

V: Teória hodnoty

264. Kde rekurzívne zdôvodňovanie narazí na dno

Prečo si myslím, že zajtra vyjde Slnko?

Pretože som videl, ako Slnko vyšlo tisíce predchádzajúcich dní.

Aha... ale prečo veríš, že budúcnosť bude ako minulosť?

Aj keby som zašiel za púhe povrchné pozorovanie vychádzajúceho Slnka, k napohľad všeobecným a bezvýnimkovým zákonom gravitácie a jadrovej fyziky, stále mi zostala otázka: „Prečo si myslíš, že toto bude pravda aj zajtra?“

Mohol by som sa odvolať na Occamovu britvu, princíp používania najjednoduchšej teórie, ktorá zodpovedá faktom... ale prečo veriť Occamovej britve? Pretože bola užitočná pri riešení problémov v minulosti? Ale kto hovorí, že to znamená, že Occamova britva bude fungovať aj zajtra?

A hľa, niekto povedal:

> „Veda je tiež založená na nepodložených predpokladoch. Preto je veda v konečnom dôsledku postavená na viere, *takže nekritizujte mňa* za to, že verím v [hlúpy-názor-číslo-238721].“

Ako som predtým poznamenal:

> Je to veľmi zvláštna psychológia – táto vec, že „Veda je tiež postavená na viere, tak vidíš!“ Zvyčajne to hovoria ľudia, ktorí tvrdia, že viera je *dobrá* vec. Prečo potom hovoria: „Veda je tiež postavená na viere!“ tým zlostným triumfálnym tónom, a nie ako kompliment?

Tvrdiť, že by ste mali byť imúnni voči kritike je zriedkavo dobrým znamením.

Ale to neodpovedaná na legitímnu filozofickú dilemu: Ak musí každý názor byť zdôvodnený, a tieto zdôvodnenia musia byť tiež zdôvodnené, ako sa potom táto nekonečná rekurzia skončí?

A ak ju môžete ukončiť niečím, čo predpokladáte bez zdôvodnenia, prečo ju potom nemôžete ukončiť *hocičím* bez zdôvodnenia?

Podobnú kritiku občas používajú proti Bayesiánstvu – že si vyžaduje predpoklad nejakých pôvodných údajov – ľudia, ktorí si asi myslia, že problém indukcie je *špeciálnym* problémom Bayesiánstva, a môžete sa mu vyhnúť použitím klasickej štatistiky.

Ale najprv tu chcem jasne priznať, že pravidlá Bayesovskej aktualizácie samotné *neriešia* problém indukcie.

Predstavte si, že ťaháte z nádoby červené a biele lopty. Pozorujete, že z prvých 9 lôpt sú 3 červené a 6 je bielych. Aká je pravdepodobnosť, že nasledujúca vytiahnutá lopta bude červená?

To závisí od vašich pôvodných predpokladov o tejto nádobe. Ak si myslíte, že si výrobca nádoby vygeneroval náhodné číslo z rovnomerného rozdelenia od 0 do 1, a použil toto číslo ako pevne danú pravdepodobnosť, že každá lopta bude červená, potom je odpoveď 4/11 (podľa Laplaceovho zákona následnosti). Ak si myslíte, že pôvodná nádoba obsahovala 10 červených a 10 bielych lôpt, potom je odpoveď 7/11.

Z čoho vyplýva, že pri správnych východiskových údajoch – alebo skôr pri nesprávnych východiskových údajoch – by sa zdalo, že pravdepodobnosť, že zajtra vyjde Slnko, každým ďalším dňom *klesá*... keby ste si boli absolútne istí, a priori, že existuje veľká nádoba, z ktorej sa každý deň losuje papierik, ktorý určuje, či Slnko vyjde alebo nie; a že táto nádoba obsahuje iba obmedzený počet papierikov, ktoré hovoria „Áno“, a použité papieriky sa nevhadzujú naspäť.

V priestore dizajnu myslí sú možné mysle, ktoré majú anti-Occamovské a anti-Laplaceovské východiská; veria, že jednoduchšie teórie majú menšiu šancu byť pravdivé, a že čím častejšie sa niečo deje, tým menšia je šanca, že sa to stane znova.

A keď sa opýtate týchto čudných bytostí, prečo stále používajú predpoklady, ktoré vyzerajú, že v skutočnom živote nikdy nefungovali... odpovedia: „Preto, lebo doteraz nikdy nefungovali!“

Takže by ste si z tohto mohli odvodiť jedno ponaučenie, a to: „Nenaroď sa s hlúpymi východiskami.“ Toto je úžasne užitočný princíp pri mnohých problémoch skutočného sveta, ale pochybujem, že by to uspokojilo filozofov.

Tu je spôsob, ako tento problém riešim ja: Snažím sa pristupovať k otázkam ako: „Mal by som dôverovať svojmu mozgu?“ alebo „Mal by som dôverovať Occamovej britve?“ ako keby to nebolo *nič špeciálne* – prinajmenšom nič špeciálne medzi hlbokými otázkami.

Mal by som dôverovať Occamovej britve? Nuž, ako veľmi to vyzerá, že Occamova britva (v hocijakej konkrétnej verzii) funguje v praxi? Aké zdôvodnenia teórie pravdepodobnosti pre ňu môžem nájsť? Keď sa pozriem na vesmír, pripadá mi ako ten typ vesmíru, kde by Occamova britva dobre fungovala?

Mal by som dôverovať svojmu mozgu? Samozrejme nie; nie vždy funguje. Ale predsa len ľudský mozog vyzerá omnoho mocnejší než ten najsofistikovanejší počítačový program, ktorý by inak prichádzal do úvahy, aby som mu dôveroval. Ako dobre funguje môj mozog v praxi, pri akom druhu problémov?

Keď preskúmam kauzálnu históriu svojho mozgu – jeho pôvod v prirodzenom výbere – nachádzam na jednej strane mnoho konkrétnych dôvodov pochybovať; môj mozog bol optimalizovaný na behanie v pravekej savane, nie na počítanie matematiky. Ale na druhej strane, je jasné prečo, voľne povedané, by tento mozog mohol aj naozaj fungovať. Prirodzený výber by rýchlo odstránil mozgy, ktoré by boli celkom *nevhodné* na rozmýšľanie, také *anti*-užitočné ako sú anti-Occamovské alebo anti-Laplaceovské východiská.

Takže to, čo som v praxi urobil, *neznamenalo* vyhlásiť náhle zastavenie pochybovania a zvôdodňovania. Nezastavujem sa v reťazi skúmania v tom bode, kde narazím na Occamovu britvu alebo svoj mozog, alebo niečo iné, o čom sa nesmie pochybovať. Reťaz pochybovania pokračuje – pokračuje však, nevyhnutne, používajúc môj terajší mozog a moje terajšie chápanie techník uvažovania. *Čo iné by som asi tak mohol použiť?*

Veru, bez ohľadu na to, *čo* by som urobil s touto dilemou, urobil by som to ja. Dokonca aj keby som dôveroval niečomu inému, napríklad nejakému počítačovému programu, bolo by to moje vlastné rozhodnutie dôverovať mu.

Technika odmietania názorov, ktoré nemajú absolútne žiadne zdôvodnenie, je vo všeobecnosti extréme dôležitá. Občas hovorím, že základnou otázkou rozumnosti je: „Prečo si myslíš to, čo si myslíš?“ Nechcel by som povedať niečo, čo by *znelo* akoby to mohlo dovoliť jedinú výnimku z pravidla, že všetko si vyžaduje zdôvodnenie.

Čo je samo osebe nebezpečná motivácia; nemôžete sa vždy vyhnúť všetkému, čo by mohlo byť riskantné, a keď vám niekto lezie na nervy hovorením niečoho hlúpeho, nemôžete obrátiť túto hlúposť a získať tak inteligenciu.

Napriek tomu by som však zdôraznil rozdiel medzi povedaním:

> „Tu je tento predpoklad, ktorý neviem zdôvodniť, ktorý sa musí jednoducho prijať a ďalej neskúmať.“

Verzus povedaním:

> „Tu bádanie pokračuje skúmaním tohto predpokladu, s plnou silou mojej *terajšej inteligencie* – čiže nie s plnou silou niečoho iného, napríklad generátora náhodných čísel alebo čarovnej veštieckej gule – aj keď sa moja terajšia inteligencia zhodou okolností zakladá na tomto predpoklade.“

Aj tak... nebolo by pekné, keby sme mohli preskúmať problém toho, ako veľmi dôverovať našim mozgom, *bez* použitia našej terajšej inteligencie? Nebolo by pekné, keby sme mohli preskúmať problém toho, ako myslieť, *bez* použitia nášho terajšieho chápania rozumnosti?

Keď to poviete *takto*, začne to vyzerať, že by odpoveď mohla byť: „Nie“.

E. T. Jaynes zvykol hovoriť, že musíte vždy použiť všetky informácie, ktoré máte k dispozícii – bol to bayesovský teoretik pravdepodobnosti a musel vyriešiť paradoxy, ktoré vytvorili iní ľudia, keď v rôznych bodoch svojich výpočtov používali rôzne informácie. Princíp: „*Vždy vynaložsvoje naozaj*

najlepšie úsilie“ má prinajmenšom toľko príťažlivosti ako: „*Nikdy nerob nič, čo by mohlo vyzerať kruhovo*.“ Napokon, alternatívou k vynaloženiu svojho najlepšieho úsilia je asi robiť menej než to najlepšie.

Ale aj tak... nebolo by pekné, keby existoval nejaký spôsob, ako zdôvodniť používanie Occamovej britvy, alebo zdôvodniť predpovedanie, že sa budúcnosť bude podobať na minulosť, *bez* predpokladania, že tie metódy uvažovania, ktoré fungovali v predchádzajúcich prípadoch, sú lepšie než tie, ktoré stále zlyhávali?

Nebolo by pekné, keby existovala nejaká reťaz zdôvodnenia, ktorá by ani nekončila nepreskúmateľným predpokladom, ani by nebola nútená skúmať sama seba podľa svojich vlastných pravidiel, ale namiesto toho by mohla byť vysvetlená od absolútneho začiatku ideálnemu študentovi filozofie dokonalej prázdnoty?

Nuž, iste by ma to zaujímalo, ale neočakávam, že také niečo v dohľadnej dobe uvidím. Už som inde na niekoľkých miestach argumentoval proti predstave, že môžete mať dokonale prázdneho ducha v stroji; neexistuje žiaden argument, ktorý môžete vysvetliť kameňu.

Dokonca aj keď niekto rozlúskne problém Prvotnej Príčiny a príde so *skutočným dôvodom, prečo je vesmír jednoduchý, ktorý samotný nebude predpokladať jednoduchý vesmír*... ešte aj tak by som očakával, že toto vysvetlenie dokáže pochopiť iba mysliaci poslucháč a nie povedzme kameň. Poslucháč, ktorý nezačal už implementujúc modus ponens, môže mať smolu.

Na záver, čo sa teda stane, keď sa ma niekto bude stále pýtať: „Prečo si myslíš to, čo si myslíš?“

V tejto chvíli sa začnem točiť do kruhu v bode, kde vysvetlím: „Predpokladám, že budúcnosť sa bude podobať na minulosť na tých najjednoduchších a najstabilnejších úrovniach organizovanosti, ktoré viem identifikovať, pretože v minulosti toto pravidlo zvyčajne úspešne vytváralo dobré výsledky; a pri použití jednoduchého predpokladu jednoduchého vesmíru môžem vidieť, *prečo* vytváralo dobré výsledky; a dokonca môžem vidieť, ako sa môj mozog mohol vyvinúť, aby pozoroval vesmír s istým stupňom presnosti, ak sú moje pozorovania správne.“

Ale potom... neodsúhlasil som práve *kruhovú logiku*?

V skutočnosti som iba odsúhlasil *reflektovanie stupňa dôveryhodnosti vašej mysle, s použitím vašej terajšej mysle namiesto niečoho iného.*

Reflexia tohto druhu je veru dôvodom, prečo v prvom rade odmietame kruhovú logiku. Chceme mať koherentný kauzálny príbeh o tom, ako naša myseľ prišla na niečo, príbeh, ktorý vysvetľuje, ako proces, ktorý sme použili na získanie našich názorov, je sám dôveryhodný. Toto je podstatná požiadavka za základnou otázkou racionalistu: „Prečo si myslíš to, čo si myslíš?“

Predstavme si teraz, že napíšete na hárok papiera: „(1) Všetko na tomto hárku je pravda. (2) Hmotnosť atómu hélia je 20 gramov.“ Keby tento trik *fungoval v skutočnom živote*, dokázali by ste vedieť skutočnú hmotnosť atómu hélia skrátka tým, že by ste verili v nejakú kruhovú logiku, ktorá by ju tvrdila. Čo by vám umožnilo nakresliť skutočnú mapu vesmíru sediac vo svojej obývačke so zavretými žalúziami. Čo by porušilo druhý zákon termodynamiky vytváraním informácie z ničoho. Čo by nebol dôveryhodný príbeh o tom, ako by vaša myseľ mohla skončiť s nejakým pravdivým názorom.

Dokonca aj keby ste začali veriť tomuto hárku papiera, nevyzeralo by to, že máte nejaký dôvod, prečo by tento papier mal zodpovedať skutočnosti. Bola by to iba zázračná zhoda okolností, že (a) hmotosť atómu hélia je 20 gramov, a (b) papier to hovorí.

Veriť sebapotvrdzujúcim množinám výrokov vo všeobecnosti nevyzerá, že by mohlo fungovať na mapovanie vonkajšej skutočnosti – keď sa nad tým *zamyslíme ako nad kauzálnym príbehom o mysli* – používajúc pritom, samozrejme, naše *terajšie* mysle.

Ale čo s vyvinutím sa tak, že viac veríme jednoduchším názorom, a veríme, že algoritmy, ktoré fungovali v minulosti, majú väčšiu šancu fungovať v budúcnosti? *Dokonca aj keď* sa nad týmto zamyslíme ako nad kauzálnym príbehom o pôvode mysle, stále to vyzerá, že by to mohlo dôveryhodne fungovať na mapovanie skutočnosti.

A čo s dôverovaním reflexívnej koherencii vo všeobecnosti? Nemohla by sa väčšina možných myslí, náhodne vytvorených a ponechaných, aby sa ustáli v stave reflexívnej koherencie, mýliť? Ach, ale my sme sa vyvinuli prirodzeným výberom; neboli sme vytvorení náhodne.

Ak vás dôverovanie tomuto argumentu znepokojuje, potom zabudnite na problém filozofického zdôvodnenia a opýtajte sa sami seba, či je to naozaj naozaj pravda.

(Pričom, samozrejme, použijete svoju vlastnú myseľ.)

Je toto to isté ako ten, kto hovorí: „Verím, že Biblia je slovo Božie, pretože to hovorí Biblia“?

Nemohli by argumentovať, že ich slepá viera do nich musela byť tiež vložená Bohom, a preto je dôveryhodná?

V skutočnosti, keď veriaci ľudia konečne prídu k odmietnutiu Biblie, *nerobia* to tak, že by čarovne skočili do neveriaceho stavu mysle čistej prázdnoty, potom by vyhodnotili svoje náboženské názory v tomto neveriacom stave mysle, a potom skočili späť do nového stavu s odstránenými náboženskými názormi.

Ľudia sa menia z veriacich sa neveriacich preto, lebo ešte aj do veriaceho stavu mysle presakuje pochybnosť. Všímajú si, že ich modlitby (a čo je horšie, modlitby napohľad omnoho hodnotnejších ľudí) zostávajú nevypočuté. Všímajú si, že Boh, ktorý k nim hovorí v ich srdci, aby im poskytol údajne uspokojujúce odpovede o vesmíre, im nedokáže povedať stú číslici čísla pí (čo by bolo omnoho upokojujúcejšie, keby Božím zámerom bolo upokojovanie). Skúmajú príbeh o Božom stvorení sveta a zatratení neveriacich, a nedáva im to zmysel ani podľa ich vlastných náboženských predpokladov.

Byť veriaci vás nerobí menej človekom. Váš mozog má stále schopnosti ľudského mozgu. Tá nebezpečná časť je, že byť veriacim vás môže zastaviť od *používania* týchto vrodených schopností na svoje náboženstvo – zastaviť vás od *plného reflektovania* seba samého. Ľudia sa neliečia zo svojich chýb tým, že by sa resetovali na ideálnych filozofov čírej prázdnoty a od začiatku by prehodnotili všetky svoje zmyslové vnemy. Liečia sa tým, že sa stávajú ochotnejšími pochybovať o svojich terajších názoroch, používajúc viac sily svojej terajšej mysle.

Preto je dôležité rozlišovať medzi *reflektovaním o svojej mysli používajúc svoju myseľ* (aj tak nemôžete použiť nič iné) a *nespochybniteľným predpokladom, o ktorom nesmiete reflektovať*.

„Verím, že Biblia je slovo Božie, pretože to hovorí Biblia.“ Nuž, keby Biblia *bola* takým neuveriteľne spoľahlivým zdrojom informácií o všetkom ostatnom, keby nehovorila, že lúčne koníky majú štyri nohy alebo že vesmír bol stvorený za šesť dní, ale namiesto toho by obsahovala periodickú tabuľku prvkov storočia pred chémiou – keby nám Biblia bola iba užitočná a hovorila by iba pravdu – potom by sme naozaj mohli mať sklon brať vážne toto dodatočné tvrdenie v Biblii, že ju vytvoril Boh. Nemohli by sme mu dôverovať úplne, lebo by to mohli byť aj mimozemšťania alebo Temní páni Matrixu, ale aspoň by sa to oplatilo brať vážne.

Podobne, keby sa všetko *ostatné*, čo nám hovorili kňazi, ukázalo ako pravda, mohli by sme brať vážnejšie ich tvrdenie, že nám vieru dal Boh a že je to systematicky dôveryhodný zdroj – najmä keby ľudia mohli zistiť aj stú číslicu čísla pí pomocou viery.

Takže tá dôležitá časť oceňovania cirkularity: „Verím, že Biblia je slovo Božie, pretože to hovorí Biblia“ nie je ani tak to, že odmietnete myšlienku reflektovania svojej mysle pomocou svojej terajšej mysle. Ale skôr to, že si uvedomíte, že všetko, čo spochybňuje dôveryhodnosť Biblie, spochybňuje aj ubezpečenie Biblie o jej vlastnej dôveryhodnosti.

Toto sa týka aj rozumnosti: keby sa budúcnosť náhodou prestala podobať na minulosť – dokonca aj na svojich najnižších a najstabilnejších pozorovaných úrovniach organizovanosti – nuž, v prvom rade by som bol mŕtvy, pretože procesy v mojom mozgu vyžadujú zákonitý vesmír, kde chémia stále funguje. Ale keby som nejako prežil, potom by som musel začať pochybovať o princípe, že by sme mali predpovedať, že sa budúcnosť bude podobať na minulosť.

Ale zatiaľ... čo je *alternatíva* k povedaniu: „Budem veriť, že sa budúcnosť bude podobať na minulosť na najstabilnejšej úrovni organizovanosti, akú viem identifikovať, pretože to mi doteraz fungovalo lepšie než hocijaký iný algoritmus, ktorý som skúsil“?

Je to povedanie: „Budem veriť, že sa budúcnosť *nebude* podobať na minulosť, pretože tento algoritmus mi doteraz vždy zlyhal“?

V tomto bode sa cítim povinný vytiahnuť argument, že racionalisti nemajú za cieľ vyhrať debaty s ideálnymi filozofmi dokonalej prázdnoty; my máme za cieľ jednoducho vyhrať. A za týmto účelom sa chceme dostať k pravde tak blízko, ako sa len dá. Takže na konci dňa sa držím princípu: „Pochybuj o svojom mozgu, pochybuj o svojich intuíciách, pochybuj o svojich princípoch rozumnosti, *s využitím celej terajšej sily svojej mysle, a v každom bode rob to najlepšie, čo môžeš*.“

Ak sa jeden z tvojich terajších princípov ukáže ako nedostatočný – podľa preskúmania tvojej vlastne mysle, keďže nemôžeš vystúpiť sám zo seba – potom ho zmeň! A potom sa vráť a pozri sa na veci znova, použijúc svoje nové vylepšené princípy.

Pointa nie je byť reflexívne konzistentným. Pointa je vyhrať. Ale *ak* sa pozeráš na seba a hráš, aby si vyhral, robíš sám seba reflexívne konzistentnejším – to je to, čo znamená „hrať, aby si vyhral“ zatiaľ čo „pozeráš sám na seba“.

Všetko, bez výnimky, potrebuje zdôvodnenie. Niekedy – pokiaľ môžem povedať, nedá sa tomu vyhnúť – tieto zdôvodnenia pôjdu dokola v reflexívnych slučkách. Myslím si, že tieto reflexívne slučky majú meta-povahu, ktorá by ich mala umožniť odlíšiť zdravým rozumom od kruhovej logiky. Ale každý, kto v prvom rade vážne uvažuje o kruhovej logike, je dokázateľne mimo vo veciach rozumnosti; a bude jednoducho trvať na tom, že jeho kruhová logika je „reflexívna slučka“ dokonca aj keď sa skladá z jediného kúska papiera s nápisom: „Dôveruj mi“. Nuž, nemôžete vždy optimalizovať svoje techniky rozumnosti na základe jediného kritéria, neumožniť tým, ktorí sa usilujú o sebazničenie, ich zneužitie.

Dôležitá vec je *nezadržiavať nič* vo svojej kritike toho, ako kritizujete; ani by ste nemali vnímať nevyhnutnosť kruhových zdôvodnení ako záruku *imunity pred spochybňovaním*.

Vždy použite plnú silu, či už obsahuje slučky alebo nie – urobte to najlepšie, čo len môžete, či už to obsahuje slučky alebo nie – a hrajte, v konečnom dôsledku, aby ste vyhrali.

* →

—

265. Môj druh reflexie

V kapitole „Kde rekurzívne zdôvodňovanie narazí na dno“ som došiel k záveru, že je okej používať indukciu na uvažovanie o pravdepodobnosti, že indukcia bude fungovať v budúcnosti, pokiaľ fungovala v minulosti; alebo používať Occamovu britvu na dôjdenie k záveru, že najjednoduchším vysvetlením, prečo Occamova britva funguje, je že samotný vesmír je v základoch jednoduchý.

Zďaleka nie som prvý človek, ktorý sa zamýšľal nad reflexívnym použitím princípov rozmýšľania. Chris Hibbert prirovnal môj pohľad k Bartleyovmu Pan-kritickému racionalizmu (bol som zvedavý, či to niekto urobí). Takže asi stojí za to, aby som ukázal, čo považujem za významné črty môjho názoru na reflexiu, čo sa môže a nemusí zhodovať s tým, čo si o reflexii mysleli iní filozofi.

- Všetky moja filozofia tu *v skutočnosti* vychádza zo snahy zistiť, ako postaviť sebamodifikujúcu UI, ktorá uplatňuje svoje vlastné princípy rozmýšľania na seba v procese prepisovania svojho vlastného zdrojového kódu. Takže kedykoľvek hovorím o použití indukcie na povolenie indukcie, *v skutočnosti* myslím na induktívnu UI, ktorá zvažuje prepísanie tej časti seba samej, ktorá robí indukciu. Ak by ste nechceli, aby UI prepísala svoj zdrojový kód tak, aby indukciu nepoužíval, potom by vaša filozofia radšej nemala označovať indukciu za nezdôvodniteľnú.
- Jeden z najsilnejších všeobecných princípov, ktoré poznám o UI vo všeobecnosti, je že sa správna Cesta vo všeobecnosti ukáže ako *naturalistická* – čo pri reflexívnom uvažovaní znamená zaobchádzať s tranzistormi vnútri UI rovnako akoby to boli tranzistory nájdené niekde v prostredí; *nie* ako ad-hoc špeciálny prípad. Toto je skutočný zdroj môjho nástojenia v „rekurzívnom zdôvodnení“, že otázky typu: „Ako dobre funguje moja verzia Occamovej britvy?“ by sa mali

→ http://lesswrong.com/lw/s0/where_recursive_justification_hits_bottom/

zvažovať rovnako ako nejaká obyčajná otázka – alebo aspoň ako obyčajná veľmi hlboká otázka. Mám silné podozrenie, že správne postavená UI pri uvažovaní, či zmeniť tú časť svojho zdrojového kódu, ktorá implementuje Occamovské uvažovanie, nebude musieť v rámci tohto zvažovania robiť nič špeciálne – konkrétne, nemala by vyvíjať špeciálne úsilie na to, aby sa vyhla používaniu Occamovského uvažovania.

• Nemyslím si, že „reflexívna koherencia“ alebo „reflexívna konzistencia“ by sa mali považovať za žiadúce sami osebe. Ako som povedal v Dvanástich cnostiach a v Jednoduchej pravde, ak nakreslíte päť presných máp toho istého mesta, potom sa tieto mapy nevyhnutne budú jedna s druhou zhodovať; ale ak nakreslíte jednu mapu podľa fantázie a potom urobíte štyri kópie, týchto päť bude zhodných, ale nie presných. Rovnako sa nikto vedome neusiluje o reflexívnu konzistenciu, a reflexívna konzistencie nie je špeciálnou zárukou dôveryhodnosti; cieľom je vyhrať. Ale každý, kto sa usiluje o cieľ vyhrávania, používajúc svoj terajší pojem vyhrávania, a upravujúc svoj vlastný zdrojový kód, nakoniec dosiahne reflexívnu konzistenciu ako vedľajší účinok – rovnako ako niekto, kto sa ustavične usiluje, aby zlepšil svoju mapu sveta, by mal nakoniec zistiť, že sa ako vedľajší účinok jednotlivé časti stávajú medzi sebou konzistentnejšie. Ak si nasadíte svoje okuliare UI, potom sa UI prepisujúca svoj vlastný zdrojový kód nepokúša urobiť sama seba „reflexívne konzistentnou“ - pokúša sa optimalizovať očakávaný úžitok svojho zdrojového kódu, a zhodou okolností to robí pomocou očakávania dôsledkov vo svojej terajšej mysli.

• Jeden zo spôsobob, ako povoľujem používanie indukcie a Occamovej britvy pri zvažovaní „indukcie“ a „Occamovej britvy“ je odvolávanie sa na princíp E. T. Jaynesa, že by sme vždy mali používať všetky dostupné informácie (ak to výpočtová sila dovoľuje) vo výpočte. Ak si myslíte, že indukcia funguje, potom by ste ju mali používať, aby ste používali svoju najväčšiu silu, vrátane toho, keď rozmýšľate o indukcii.

• Vo všeobecnosti si myslím, že sa oplatí rozlišovať medzi obranným postojom, keď si predstavujete, ako by ste zdôvodnili svoju filozofiu pred filozofom, ktorý vás spochybňuje, a útočným postojom, keď sa snažíte dostať tak blízko k pravde, ako sa len dá. Nejde teda o to, že mať podozrenie ohľadom Occamovej britvy, ale použiť svoju terajšiu myseľ a inteligenciu na jej preskúmanie ukazuje, že ste *féroví* a môžete sa *hájiť*, že pochybujete o svojich základných princípoch. Ale skôr, že dôvod, prečo skúmate Occamovu britvu je, aby ste videli, či viete vylepšiť, ako ju používate, alebo že sa obávate, že by v skutočnosti mohla byť nesprávna. Mám sklon odsudzovať pochybnosti z púhej povinnosti.

• Ak pracujete na skúmaní svojich základov, očakávam od vás, že ich naozaj zlepšíte, nie iba povinne preskúmate. Naše mozgy sú postavené tak, aby odhadovali „jednoduchosť“ istým intuitívnym spôsobom, pri ktorom Thor znie jednoduchšie než Maxwellove rovnice ako vysvetlenie blesku. Keď už však máme lepší pohľad na to, ako vesmír naozaj funguje, došli sme k záveru, že diferenciálne rovnice (ktoré zvláda málo ľudí) sú v skutočnosti *jednoduchšie* (v zmysle teórie informácií) než hrdinská mytológia (ktorou väčšina kmeňov vysvetľuje vesmír). Keďže je to takto, snažili sme sa aj našu predstavu Occamovej britvy preložiť do matematiky.

• Na druhej strane, tieto vylepšené základy by stále mali dávať dokopy normálnosť; 2 + 2 by sa stále malo nakoniec rovnať 4, nie niečomu novému a úžasnému a vzrušujúcemu, ako „ryba“.

• Myslím si, že je veľmi dôležité rozlišovať medzi otázkami: „Prečo funguje indukcia?“ a „Funguje indukcia?“ Dôvod, *prečo je samotný vesmír pravidelný*, je pre nás stále tajomnou otázkou, zatiaľ. Zvláštne špekulácie tu môžu byť dočasne nevyhnutné. Ale na druhej strane, ak začnete tvrdiť, že vesmír *v skutočnosti nie je pravidelný*, že odpoveď na otázku: „Funguje indukcia?“ je „Nie!“, potom sa dostávate do oblasti 2 + 2 = 3. Snažíte sa urobiť svoju filozofiu veľmi zaujímavou, namiesto správnou. Induktívna UI, ktorá sa pýta, aký odhad pravdepodobnosti urobiť v nasledujúcom kole, sa pýta: „*Funguje* indukcia?“ a toto je otázka, na ktorú môže odpovedať pomocou induktívneho uvažovania. Ak sa pýtate: „*Prečo* funguje indukcia?“, potom

odpoveď: „Lebo indukcia funguje“ je kruhová logika a odpoveď: „Lebo verím, že indukcia funguje“ je magické myslenie.

- Nemyslím si, že ísť dokola v slučke zdôvodnení na meta-úrovni je to isté ako kruhová logika. Myslím si, že pojem „kruhová logika“ sa týka objektovej úrovne, a je to niečo definitívne zlé a zakázané, na objektovej úrovni. Zakazovať *reflexívnu koherenciu* neznie ako dobrý nápad. Ale zatiaľ som si nesadol a nesformalizoval presný rozdiel – moja reflexívna teória je niečo, čo sa snažím vypracovať, nie niečo, čo už mám v hrsti.

⁂→

266. Neexistujú univerzálne presvedčivé argumenty

Čo je také *desivé* na predstave, že nie každá možná myseľ s nami môže z princípu súhlasiť?

Pre niektorých ľudí, nič – ani v najmenšom ich to netrápi. A pre niektorých z *týchto* ľudí je *dôvodom*, prečo ich to netrápi, nedostatok silných intuícií o štandardoch a pravdách, ktoré prekračujú osobné rozmary. Ak povedia, že obloha je modrá, alebo že vražda je zlá, je to iba ich osobný názor; a neprekvapí ich, že niekto iný by mohol mať iný osobný názor.

Pre iných ľudí, nesúhlas, ktorý *z princípu* pretrváva, je niečo neprijateľné. A pre niektorých z *týchto* ľudí je *dôvodom*, prečo ich to trápi, dojem, že ak pripustíte, že niektorých ľudí *z princípu* nemožno presvedčiť, že obloha je modrá, potom priznávate, že „obloha je modrá“ je iba *svojvoľný* osobný názor.

Navrhol som, že by ste mali odolať pokušeniu zovšeobecňovať pre celý priestor dizajnu myslí. Ak sa obmedzíme na mysle, ktoré možno určiť nanajvýš bilión bitmi, potom každé *všeobecné* zovšeobecnenie „Pre každú myseľ m: X(m)“ má šancu dve na bilióntu byť nepravdivé, zatiaľ čo každé *existenčné* zovšeobecnenie „Existuje myseľ m: X(m)“ má šancu dve na bilióntu byť pravdivé.

Toto by napohľad znamenalo, že pre každý argument A, nech už sa nám zdá akokoľvek presvedčivý, existuje aspoň jedna možno myseľ, ktorá to neprijme.

A to prekvapenie a/alebo hrôza z tejto predstavy (pre niektorých) podľa mňa do veľkej miery súvisí s intuíciou ducha v stroji – ducha s nejakým ireducibilným jadrom, ktorého presvedčí ľubovoľný *naozaj správny* argument.

Už som hovoril o intuícii, podľa ktorej si ľudia mapujú *programovanie počítača* na *dávanie príkazov ľudskému sluhovi*, takže by sa počítač mohol vzbúriť proti svojmu kódu – alebo možno pozrieť na svoj kód, rozhodnúť sa, že je nerozumný, a vrátiť ho naspäť.

Keby existoval duch v stroji, a keby tento duch obsahoval neredukovateľné jadro rozumnosti, voči ktorému by hocijaký kód bol iba odporučením, potom by mohli existovať univerzálne argumenty. Dokonca aj keby tento duch pôvodne dostal kód-odporučenie v rozpore s Univerzálnym Argumentom, keby sme duchovi neskôr ukázali tento Univerzálny Argument – alebo keby stroj objavil tento Univerzálny Argument sám, to je tiež obľúbená predstava – potom by duch jednoducho prehlasoval svoj vlastný pomýlený zdrojový kód.

Ale ako raz povedal istý študent programovania: „Mám taký pocit, akoby počítač všetky tieto komentáre jednoducho preskakoval.“ Ten kód nie je niečo, čo dáme UI; ten kód *je* UI.

Ak sa prepnete na fyzikálny pohľad, potom predstava Univerzálneho Argumentu vyzerá zrejme nefyzikálne. Ak existuje fyzikálny systém, ktorý v čase T, po vystavení argumentu E, urobí X, potom by mal existovať iný fyzikálny systém, ktorý v čase T, po vystavení argumentu E, urobí Y. Ľubovoľná myšlienka musí byť implementované *niekde* vo fyzikálnom systéme; ľubovoľný názor, ľubovoľné rozhodnutie, ľubovoľný motorický výstup. Pre každý zákonitý kauzálny systém, ktorý pri nejakej množine bodov urobí cik, by ste mali vedieť špecifikovať iný kauzálny systém, ktorý pri tých istých bodoch zákonite urobí cak.

Povedzme, že existuje myseľ, ktorej tranzistor dá v čase T výstup +3 volty, čo znamená, že práve súhlasila s nejakým presvedčivým argumentom. Potom môžeme postaviť vysoko podobný fyzikálny

→ http://lesswrong.com/lw/s2/my_kind_of_reflection/

kognitívny systém a maličkým prepadliskom pod týmto tranzistorom, v ktorom bude malý sivý mužíček, ktorý v čase T vylezie a nastaví výstup tohto tranzistora na -3 volty, čo znamená nesúhlas. Na tom nie je nič nekauzálne; ten malý sivý mužíček je tam preto, lebo sme ho tam zabudovali. Predstava argumentu, ktorý presvedčí *ľubovoľnú* myseľ, asi zahŕňa malú modrú žienku, ktorú do systému *nikto nikdy* nezabudoval, ktorá vylezie doslova *odnikiaľ* a zaškrtí tohto malého sivého mužíčka, pretože daný tranzistor skrátka *musí* dať výstup +3 volty: Ako vidíte, je to taký *presvedčivý argument*.

Lenže presvedčivosť nie je vlastnosťou argumentu, je to vlastnosť mysle, ktorá argument spracováva.

Takže dôvod, prečo argumentujem proti duchovi, nie je *iba* aby som povedal, že (1) Priateľská UI musí byť explicitne naprogramovaná, a že (2) fyzikálne zákony nezakazujú Priateľskú UI. (Hoci samozrejme mám istý záujem na objasnení tohto.)

Chcem aj objasniť predstavu mysle ako *kauzálneho, zákonitého, fyzikálneho systému,* v ktorom neexistuje *žiaden* neredukovateľný duch, ktorý dozerá na neuróny / kód a rozhoduje sa, či sú to dobré odporúčania.

(V oblasti Priateľskej UI existuje predstava, že *úmyselne* naprogramujeme PUI, aby skontrolovala svoj vlastný zdrojový kód a prípadne ho vrátila programátorom. Lenže tá myseľ, ktorá kontroluje, nie je nereducibilná, je to myseľ, ktorú ste práve vytvorili. PUI sa renormalizuje *tak, ako ste ju nadizajnovali, aby to robila*; nie je tam nič nekauzálne, zasahujúce zvonka. Je to lešenie, nie žeriav.)

Toto všetko odráža obavu z bayesovských „ľubovoľných“ pôvodných údajov. Ak mi ukážete Bayesiána, ktorý vytiahne zo suda 4 červené gule a 1 bielu, a ktorý priradí vytiahnutiu červenej gule v nasledujúcom ťahu pravdepodobnosť 5/7 (podľa Laplaceovho pravidla pokračovania), ja vám viem ukázať inú myseľ, ktorá sa riadi Bayesovým pravidlom a ktoré vytiahnutiu červenej gule v nasledujúcom ťahu priradí pravdepodobnosť 2/7 – čo zodpovedá tomu, že má o sude inú pôvodnú predstavu, hoci možno menej „rozumnú“.

Mnohí filozofi sú presvedčení, že keďže môžete v princípe vytvoriť pôvodný predpoklad, ktorý sa po prúde indícií aktualizuje na ľubovoľný záver, bayesovské uvažvanie musí byť preto „svojvoľné“, a celá schéma bayesiánstva je pomýlená, pretože spočíva na „nezdôvodniteľných“ predpokladoch, a dokonca až „nevedeckých“, pretože nedokážete donútiť ľubovoľného možného redaktora odborného časopisu v priestore myslí, aby s vami súhlasil.

Lenže toto (odpovedal som), stojí na predpoklade, že odmotaním všetkých argumentov a ich zdôvodnení dokážete získať ideálneho študenta filozofie dokonalej prázdnoty, ktorého možno presvedčiť líniou tvrdení nevychádzajúc z absolútne žiadnych predpokladov.

Lenže kto je tento ideálny filozof absolútnej prázdnoty? Nuž, to je práve to neredukovateľné jadro toho ducha!

A to je dôvod, prečo (hovorím ďalej) výsledkom snahy odstrániť z mysle všetky predpoklady, a rozmotať ju k dokonalej neprítomnosti ľubovoľného východiska, nie je ideálny filozof dokonalej prázdnoty, ale kameň. Čo vám zostane z mysle, keď odstránite zdrojový kód? Nie ten duch, ktorý dozeral na zdrojový kód, ale iba... bezduchá prázdnota.

Takže – a k tejto téme sa ešte neskôr vrátim – kamkoľvek umiestňujete vaše predstavy o *správnosti* alebo *hodnote* alebo *rozumnosti* alebo *zdôvodnení* alebo dokonca *objektivite*, nemôžete to oprieť o argument, ktorý je *univerzálne presvedčivý pre všetky fyzikálne možné mysle.*

Ani nemôžete založiť platnosť na postupnosti zdôvodnení takých, že by začínajúc z ničoho presvedčili dokonalú prázdnotu.

Áno, možno existuje postupnosť argumentov, ktoré by primäli ľubovoľného neurologicky nepoškodeného *človeka* – ako argument, ktorý používam na ľudí, aby vypustili UI z krabice[243] – ale to z filozofického pohľadu vôbec nie je to isté.

Prvá veľká chyba tých, ktorí si skúšajú predstaviť Priateľskú UI, je Jeden Veľký Morálny Princíp, Ktorý Jediný Potrebujeme Naprogramovať – alias falošná funkcia úžitku – a o tom som už hovoril.

243 Len žartujem.

Ale ešte horšia chyba je Jeden Veľký Morálny Princíp, Ktorý Dokonca Ani *Nepotrebujeme* Naprogramovať, Pretože Každá UI Ho Nevyhnutne Objaví. Táto predstava vyvoláva desivo nezdravú fascináciu u tých, ktorí ju spontánne znovuobjavia; snívajú o príkazoch, ktoré by žiadna dostatočne vyspelá myseľ nedokázala neposlúchnuť. Samotní bohovia potvrdia správnosť ich filozofie. (Napríklad John C. Wright, Marc Geddes.)

Existuje aj menej vážna verzia tejto chyby, kde človek *nevyhlasuje* Jedinú Pravú Morálku. Namiesto toho dúfa v UI vytvorenú ako *dokonale slobodná*, neobmedzená skazenými ľuďmi želajúcimi si otroka, aby táto UI mohla objaviť cnosť podľa svojej vôle – takú cnosť, o ktorej možno ani nesníval hovoriaci, ktorý sa priznáva, že je príliš skazený, než aby učil UI. (Napríklad John K Clark, Richard Hollerith?, Eliezer$_{1996}$.) Toto je menej skazený motív než túžba po absolútnom príkaze. Ale aj keď *tento* sen vychádza skôr z cnosti než z neresti, stále je založený na nesprávnom pochopení slobody, a *v skutočnom živote* nebude naozaj *fungovať*. O tomto, samozrejme, bude ešte viac.

John C. Wright, ktorý pôvodne písal veľmi peknú transhumanistickú trilógiu (prvá kniha: *Zlatý vek*), vložil doprostred svojej záverečnej tretej knihy obrovskú Autorskú Prednášku, opisujúcu na desiatkach stránok jeho Univerzálnu Morálku, Ktorá Musí Presvedčiť Ľubovoľnú UI. Neviem, či sa potom niečo stalo, pretože tam som prestal čítať. A potom Wright konvertoval na kresťanstvo – áno, vážne. Takže do tejto pasce *naozaj nechcete* spadnúť.

*→

267. Už stvorení v pohybe

Lewis Carroll, ktorý bol zároveň matematikom, raz napísal krátky dialóg s názvom Čo Korytnačka povedala Achilovi. A ste ešte nečítali túto starú klasiku, skúste to urobiť teraz.

Korytnačka ponúka Achilovi krok uvažovania podľa Euklidovho Prvého tvrdenia:

> (A) Veci, ktoré sa rovnajú tomu istému, sa rovnajú navzájom.
>
> (B) Dve strany tohto trojuholníka sa rovnajú tomu istému.
>
> (Z) Dve strany tohto trojuholníka sa rovnajú navzájom.

Korytnačka: „A ak nejaký čitateľ ešte *neprijal* A a B ako pravdivé, stále by asi mohol prijať túto *postupnosť* ako *správnu*, predpokladám?“

Achilles: „Nepochybujem, že by taký čitateľ mohol existovať. Mohol by povedať: ‚Prijímam ako pravdivé hypotetické tvrdenie, že *ak* sú A a B pravdivé, Z musí byť pravdivé; ale *neprijímam* A a B ako pravdivé.‘ Taký čitateľ by urobil múdro, keby zanechal Euklida a šiel hrať futbal.“

Korytnačka: „A nemohol by existovať *aj* nejaký čitateľ, ktorý by povedal: ‚Prijímam A a B ako pravdivé, ale *neprijímam* záver?‘“

Achiles v tomto nemúdro ustúpi; a tak požiada Korytnačku, aby prijala ďalší výrok:

> (C) Ak A a B sú pravdivé, Z musí byť pravdivé.

Ale, opýta sa Korytnačka, čo ak prijme A a B a C, ale nie Z?

Potom, povie Achiles, musí požiadať Korytnačku, aby prijala ešte jeden predpoklad:

> (D) Ak A a B a C sú pravdivé, Z musí byť pravdivé.

Douglas Hofstadter parafrázoval tento argument niekedy neskôr:

> Achiles: Ak máš [(A a B) → Z], a máš aj (A a B), potom iste máš aj Z.
>
> Korytnačka: Aha! Myslíš <{(A a B) a [(A a B) → Z]} → Z>, však?

Ako hovorí Hofstadter: „Čokoľvek Achiles považuje za pravidlo odvodenia, Korytnačka to ihneď sploští na púhy reťazec v systéme. Ak použijete iba písmená A, B a Z, dostanete rekurzívny vzor dlhších a dlhších reťazcov.“

→ http://lesswrong.com/lw/rn/no_universally_compelling_arguments/

Teraz už by ste mali rozoznať anti-vzor Odovzdávanie rekurzívnej mince; a hoci sa protiliek niekedy ťažko hľadá, keď sa nájde, vo všeobecnosti má tvar Minca Zastane Ihneď.

Myseľ Korytnačky potrebuje *dynamiku* pridania Y do množiny presvedčení, keď X a (X → Y) už sú v množine presvedčení. Ak táto dynamika neexistuje – napríklad kameň ju nemá – potom môžete pridávať X a (X → Y) a (X a (X→Y)) → Y až do konca sveta, a nikdy sa nedostanete k Y.

Raz mi pri popise tejto požiadavky napadlo slovné spojenie, že myseľ musí byť *už stvorená v pohybe*. Neexistuje žiaden argument natoľko presvedčivý, aby dal túto dynamiku statickej veci. Neexistuje žiaden počítačový program taký *presvedčivý*, že ho môžete zbehnúť na kameni.

A dokonca aj keby ste mali myseľ, ktoré *dokáže* vykonať modus ponens, je márne, aby mala názory ako...

> (A) Ak je na vlakových koľajach batoľa, potom odtiahnuť ho preč je fuzzle.
>
> (B) Na vlakových koľajach je batoľa.

...pokiaľ táto myseľ zároveň *neimplementuje*:

> *Dynamika:* Ak množina názorov obsahuje „X je fuzzle“, pošli X pohybovému systému.

Slovom „dynamika“ myslím vlastnosť *vývoja* fyzikálne implementovaného kognitívneho systému *v čase*. „Dynamika“ je niečo, čo sa *stane vnútri* kognitívneho systému, *nie* údaje, ktoré skladuje v pamäti a manipuluje s nimi. Dynamika sú tie manipulácie. Neexistuje spôsob, ako napísať dynamiku na kus papiera, pretože ten papier tam bude iba ležať. Takže horeuvedený text, ktorý hovorí „dynamika“ nie je dynamika. Keby som chcel, aby ten text *bol* dynamika, a nie iba *hovoril* „dynamika“, musel by som napísať Java applet.

Netreba dodávať, že mať názor...

> (C) Ak množina názorov obsahuje „X je fuzzle“, potom „pošli ‚X‘ pohybovému systému“ je fuzzle.

...nepomôže, pokiaľ myseľ už neimplementuje *správanie* prekladania hypotetických akcií označených ako „fuzzle“ do skutočných pohybových akcií.

Pomocou starostlivého argumentovania o povahe kognitívnych systémov by sa vám možno podarilo dokázať...

> (D) Myseľ s dynamikou, ktorá posiela plány označené ako „fuzzle“ do pohybového systému, je viac fuzzle než mysle, ktoré to nerobia.

...ale to by vám *stále* nepomohlo, pokiaľ by počúvajúca myseľ už *predtým* nemala *dynamiku* nahradenia svojho zdrojového kódu alternatívnym zdrojovým kódom, o ktorom by verila, že je viac fuzzle.

To je dôvod, prečo nemôžete ukecať kameň, aby bol fuzzle.

* →

268. Triedenie kamienkov na správne kôpky

Kde bolo, tam bolo, bol raz jeden zvláštny živočíšny druh – ktorý bol možno biologický, alebo možno syntetický, alebo to možno bol iba sen – ktorého vášňou bolo triediť kamienky na správne kôpky.

Nevedeli by vám povedať, *prečo* sú nejaké kôpky správne a iné nesprávne. Ale všetci sa zhodli na tom, že najdôležitejšou vecou na svete je tvoriť správne kôpky, a rozhadzovať tie nesprávne.

Prečo na tom Triedičom tak veľmi záležalo, je stratené v histórii – možno to bol splašený Fisherovský sexuálny výber, naštartovaný čírou náhodou pred miliónom rokov? Alebo možno zvlášne dielo vedomého umenia, vytvorené mocnejšou mysľou a opustené?

Ale ukrutne im na tom záležalo, na tomto triedení kamienkov, takže všetci Triediči filozofi jednohlasne hovorili, že triedenie kamienkov na kôpky je najvyšším zmyslom ich životov: a trvali na tom,

→ http://lesswrong.com/lw/rs/created_already_in_motion/

že jediným oprávneným dôvodom, prečo jesť, je aby sme triedili kamienky, jediným oprávneným dôvodom, prečo sa páriť, je aby sme triedili kamienky, jediným oprávneným dôvodom účasti na ekonomike ich sveta je, aby sme efektívnejšie triedili kamienky.

Triediči sa všetci zhodli na tom, ale nie vždy sa zhodli na tom, ktoré kôpky sú správne alebo nesprávne.

V raných dňoch civilizácie Triedičov boli vyrábané kôpky prevažne malé, s počtami ako 23 alebo 29; nevedeli povedať, či sú väčšie kôpky správne alebo nie. Pred troma tisícročiami Veľký Vodca Biko urobil kôpku s 91 kamienkami a vyhlásil ju za správnu, a jeho hordy obdivujúcich nasledovateľov urobila veľa takýchto kôpiek. Ale počas niekoľkých storočí, ako moc Bikovcov upadala, začali tí najbystrejší a najvzdelanejší získavať dojem, že kôpka s 91 kamienkami je nesprávna. Až napokon zistili, čoho sa dopustili: a potom rozhádzali všetky tie kôpky z 91 kamienkov. Nebolo to bez záblesku ľútosti, pretože niektoré z tých kôpiek boli veľké umelecké diela, ale nesprávne. Rozhádzali dokonca aj Bikovu pôvodnú kopu, vytvorenú z 91 vzácnych drahokamov, každý iného druhu a farby.

A odvtedy žiadna civilizácia vážne nepochybovala, že kôpka 91 je nesprávna.

Dnes, v týchto múdrejších dobách, veľkosť kôpiek, o ktoré sa Triediči opovážia pokúsiť, výrazne vzrástla – na čom sa všetci zhodnú, že by bola tá najúžasnejšia a najskvelejšia vec, keby len vedeli zabezpečiť, že tie kôpky sú naozaj *správne*. Krajiny, ktoré nesúhlasili, ktoré kôpky sú správne, navzájom bojovali: Triediči nikdy nezabudnú na Veľkú Vojnu o 1957, ktorú medzi sebou viedli Y'ha-nthlei a Y'not'ha-nthlei o kôpky veľkosti 1957. Túto vojnu, kde sa na planéte Triedičov po prvýkrát použili jadrové zbrane, ukončil až Y'not'ha-nthleiský filozof At'gra'len'ley ukážkou kôpky s 103 kamienkami a kôpky s 19 kamienkami vedľa seba. Tento argument bol taký presvedčivý, že dokonca aj Y'not'ha-nthlei neochotne pripustili, že by bolo lepšie prestať stavať kôpky s 1957 kamienkami, prinajmenšom dočasne.

Od Veľkej Vojny o 1957 sa krajiny vyvarovali otvoreného schvaľovania alebo zakazovania kôpiek veľkých rozmerov, lebo to tak ľahko viedlo k vojne. Dokonca niektorí Triediči filozofi – ktorí vyzerajú, že majú hmatateľné potešenie zo šokovania druhých svojim cynizmom – celkom popierali existenciu *pokroku* v triedení kamienkov; naznačovali, že názory o kamienkoch sa jednoducho časom vyvíjajú náhodným smerom, nie je v nich žiadna koherencia, a ilúzia pokroku vzniká odsudzovaním všetkých odlišných vecí v minulosti ako nesprávnych. Títo filozofi poukazovali na nesúhlas ohľadom veľkých kôpiek, ako na dôkaz, že v skutočnosti nič nerobí kôpku veľkosti 91 naozaj *nesprávnou* – že bolo jednoducho v istej dobe módou stavať takéto kôpky, a potom v inej dobe módou preklínať ich. „Ale... 13!“ na nich nijako nepôsobilo; pretože vnímať „13!“ ako presvedčivý protiargument je iba ďalším zvykom, tvrdili. Kôpkoví Relativisti tvrdili, že ich filozofia môže zabrániť budúcim katastrofám ako bola Veľká Vojna o 1957, ale boli prevažne považovaní za filozofiu zúfalstva.

Teraz sa otázka, čo robí kôpku správnou alebo nesprávnou, stala znovu naliehavou; pretože Triediči môžu čoskoro prísť k vytvoreniu sebazdokonaľujúcich sa Umelých Inteligencií. Kôpkoví Relativisti pred týmto projektom varovali: Hovorili, že UI, ktoré nebudú patriť k druhu *Triedič sapiens*, si môžu vytvoriť svoju vlastnú kultúru s celkom odlišnými predstavami o tom, ktoré kôpky sú správne alebo nesprávne. „Mohli by sa rozhodnúť, že kôpky s 8 kamienkami sú správne,“ hovorili Kôpkoví Relativisti, „a hoci by v konečnom dôsledku neboli o nič správnejší ani nesprávnejší než my, predsa len, *naša* civilizácia hovorí, že by sme takéto kôpky nemali stavať. Nie je v našom záujme vytvoriť UI, dokiaľ všetky počítače nebudú mať k sebe pripevnené bomby, takže keby si UI čo len pomyslela, že kôpka z 8 kamienkov je správna, môžeme ju donútiť namiesto toho stavať kôpky zo 7 kamienkov. Inak, BUM!“

Lenže toto väčšine Triedičov pripadalo absurdné. Nepochybne by dostatočne mocná UI – najmä taká „superinteligencia“, o ktorej niektorí transtriediči stále rozprávajú – dokázala rozoznať *na prvý pohľad*, ktoré kôpky sú správne a nesprávne! Predstava, že by si niečo s mozgom veľkosti planéty mohlo myslieť, že kôpka z 8 kamienkov je správna, je skrátka príliš absurdná, než aby sa o tom oplatilo hovoriť.

A vôbec, je vrcholne márne plánovať, ako obmedziť superinteligenciu pri triedení kamienkov na kôpky. Predstavte si, že by Veľký Vodca Biko dokázal vo svojej primitívnej dobe zostrojiť sebazdokonaľujúcu sa UI; a že by ju postavil ako maximalizátor úžitku, ktorého funkcia úžitku by jej hovorila, aby postavila čo najviac kôpiek veľkosti 91. Iste, keby sa táto UI dostatočne zdokonalila, a stala

by sa dosť bystrou, dokázala by napohľad rozoznať, že jej funkcia úžitku je nesprávna; a keďže by mala schopnosť prepísať svoj zdrojový kód, *prepísala by svoju funkciu úžitku*, aby si cenila rozumnejšie veľkosti kôpiek, napríklad 101 alebo 103.

A určite nie kôpky veľkosti 8. To by bolo jednoducho *hlúpe*. Myseľ, ktorá je natoľko hlúpa, je príliš hlúpa na to, aby bola hrozbou.

Povzbudení toľkým zdravým rozumom sa Triediči vrhli plnou parou vpred do svojho projektu spájania algoritmov náhodnýcm spôsobom na veľkých počítačoch, kým sa nevynorí nejaký druh inteligencie. Celé dejiny civilizácie ukazovali, že bohatšie, bystrejšie, lepšie vzdelané civilizácie sa s väčšou pravdepodobnosťou zhodnú na kôpkach, o ktorých sa ich predkovia kedysi hádali. Iste, potom sa dá hádať o ešte väčších kôpkach – ale ako sa technológia postupne zdokonaľovala, dalo sa dohodnúť na väčších kôpkach a postaviť ich.

Napokon, samotná inteligencia vždy korelovala s robením kôpiek správnej veľkosti – najbližší evoluční príbuzní Triedičov, Triedimpanzi, robili kôpky veľkosti iba 2 alebo 3, a občas hlúpe kôpky ako 9. A iné, ešte menej inteligentné tvory, napríklad ryby, nerobili vôbec žiadne kôpky.

Múdrejšie mysle rovná sa múdrejšie kôpky. Prečo sa tento trend mal prerušiť?

*→

269. 2-miestne a 1-miestne slová

Už som hovoril o dávnych obálkach brakových časopisov, ktoré znázorňovali príšeru s hmyzími očami, ako odnáša devu v roztrhaných šatoch; a ako si ľudia myslia, že príťažlivosť je vnútorná vlastnosť sexi bytosti, bez ohľadu na obdivovateľa.

„*Samozrejme*, že by príšera s modrými očami uprednostňovala ľudské ženy pred svojím vlastným druhom," povie umelec (nazvime ho Fred); „vidí predsa, že ľudské ženy majú príjemnú, mäkkú pokožku namiesto slizkých šupín. Môže to byť mimozemšťan, ale nie je to *blbec* – prečo si myslíte, že by urobil takú základnú chybu ohľadom príťažlivosti?"

Čo je Fredova chyba? Je to zaobchádzanie s funkciou, ktorá má 2 argumenty („2-miestnou funkciou"):

Príťažlivosť: Obdivovateľ, Bytosť -> [0; ∞)

akoby to bola funkcia s 1 argumentom („1-miestna funkcia"):

Príťažlivosť: Bytosť -> [0; ∞)

Ak berieme Príťažlivosť ako funkciu, ktorá prijíma iba jednu Bytosť ako svoj argument, potom sa samozrejme bude zdať, že Príťažlivosť závisí iba od Bytosti, a nič iné nie je podstatné.

Keď myslíte na dvojmiestnu funkciu akoby to bola jednomiestna funkcia, skončíte s Klamom otázky s premennou / Klamom projekcie mysle. Ako keď sa pokúšate určiť, či je budova *objektívne* na ľavej alebo na pravej strane cesty, bez ohľadu na to, akým smerom kto cestuje.

Alternatívne a rovnako platné stanovisko je, že „príťažlivosť" *odkazuje* na jednomiestnu funkciu – ale každý hovoriaci používa *inú* jednomiestnu funkciu na rozhodovanie, koho by uniesol a zneuctil. Kto hovorí, že keď umelec Fred a príšera s hmyzími očami Blúga obaja používajú slovo „sexi", musia tým myslieť tú istú vec?

Ak prijmeme toto stanovisko, nie je žiaden paradox v tom, keď povieme, že nejaká žena objektívne má 5 jednotiek na Fred::Príťažlivosť. Všetci pozorovatelia sa na tomto fakte môžu zhodnúť, akonáhle sa špecifikuje Fred::Príťažlivosť v pojmoch kriviek, textúry pokožky, oblečenia, náznakov postavenia, atď. Táto špecifikácie *nemusí vôbec spomínať Freda*, iba hodnotenú ženu.

Zhodou okolností Fred *používa* tento algoritmus na výber cieľov flirtovania. Ale to neznamená, že samotný algoritmus musí *spomínať* Freda. Fredova funkcia Príťažlivosť z tohto pohľadu naozaj *je*

→ http://lesswrong.com/lw/sy/sorting_pebbles_into_correct_heaps/

funkciou jedného objektu – danej ženy. Nazývam to Fred::Príťažlivosť, ale pamätajte, že toto *meno* odkazuje na funkciu, ktorá je opísaná nezávisle od Freda. Možno by bolo lepšie napísať:

Fred::Príťažlivosť = Príťažlivosť_20934

Je empirickým faktom o Fredovi, že používa funkciu Príťažlivosť_20934 na vyhodnocovanie potenciálnych partneriek. Možno John používa presne ten istý algoritmus; nezáleží na tom, odkiaľ pochádza, keď ho už raz máme.

A podobne, tá istá žena má iba 0,01 jednotiek na Príťažlivosť_72546, pričom slizká forma má 3 jednotky príťažlivosti na Príťažlivosť_72546. Zhodou okolností je empirický fakt, že Blúga používa Príťažlivosť_72546 na rozhodovanie, koho uniesť; čiže Blúga::Príťažlivosť označuje matematický objekt nezávislý na Blúge, ktorým je funkcia Príťažlivosť_72546.

Keď povieme, že daná žema má 0,01 jednotiek na Príťažlivosť_72546 a 5 jednotiek na Príťažlivosť_20934, všetci pozorovatelia sa na tom dokážu zhodnúť bez paradoxov.

A tak možno 2-miestny a 1-miestny pohľad zjednotiť pomocou pojmu „currying“, pomenovaného podľa matematika Haskelly Curryho. Currying je technika povolená v niektorých programovacích jazykoch, kde napríklad namiesto písania

x = plus(2, 3) (x = 5)

môžete aj napísať

y = plus(2)

(y je teraz „curryovaná“ forma funkcie plus, ktorá zjedla číslo 2)

x = y(3) (x = 5)

z = y(7) (z = 9)

Takže plus je 2-miestna funkcia, ale curryovanie plus – nechanie ho zjesť iba jeden zo svojich dvoch požadovaných argumentov – z neho urobí 1-miestnu funkciu, ktorá pripočíta 2 k ľubovoľnému vstupu. (Podobne môžete začať so 7-miestnou funkciou, nakŕmiť do nej 4 argumenty, a výsledkom bude 3-miestna funkcia, atď.)

Skutočný purista by trval na tom, že každú funkciu treba vidieť, podľa definície, ako funkciu s presne 1 argumentom. Z tohto pohľadu plus akceptuje 1 číselný vstup, a výstupom je *nová* funkcia; a táto *nová* funkcia má 1 číselný vstup a výstupom konečne je číslo. Z tohto pohľadu, keď napíšeme plus(2, 3), v skutočnosti počítame plus(2), aby sme dostali funkciu, ktorá pripočíta 2 k ľubovoľnému vstupu, a potom aplikujeme výsledok na 3. Programátor by to zapísal takto:

plus: int -> (int -> int)

Toto hovorí, že plus berie int ako argument, a vráti funkciu typu int -> int.

Ak túto metaforu preložíme späť do ľudských slov, mohli by sme si predstaviť, že „príťažlivost“ začne tým, že pohltí objekt Obdivovateľ a vypľuje pevne daný *matematický* objekt, ktorý opisuje, ako daný Obdivovateľ vyhodnocuje nádheru. Je to *empirický* fakt o danom Obdivovateľovi, že jeho intuície o túžbe sa počítajú spôsobom, ktorý je izomorfný s touto *matematickou* funkciou.

Potom ten matematický objekt získaný curryovaním Príťažlivosť(Obdivovateľ) možno aplikovať na objekt Žena. Ak pôvodný Obdivovateľ bol Fred, Príťažlivosť(Fred) najprv vráti Príťažlivosť_20934. Potom môžeme povedať, že je empirický fakt o danej *Žene*, nezávisle na Fredovi, že Príťažlivosť_20934(Žena) = 5.

V myšlienkovom experimente Hilaryho Putnama o „dvojčati Zeme“ bol ohromný filozofický rozruch, či má zmysel predpokladať Dvojča Zeme, ktoré je rovnaké ako naša Zem, akurát že voda nie je H20, ale nejaká *odlišná* priehľadá tečúca látka XYZ. Navyše, nastavme čas tohto myšlienkového experimentu pár stotočí dozadu, takže na našej Zemi ani na Dvojčati Zeme nikto nevie ako otestovať alternatívne hypotézy H2O verzus XYZ. *Znamená* slovo „voda“ v takom svete to isté ako v našom?

Niektorí povedali: „Áno, pretože keď človek zo Zeme a človek z Dvojčaťa Zeme vyslovia slovo ‚voda‘, majú na mysli rovnaké zmyslové testy.“

Niektorí povedali: „Nie, pretože ‚voda' na našej Zemi znamená H20 a ‚voda' na Dvojčati Zeme znamená XYZ."

Ak si myslíte, že „voda" je pojem, ktorý *začne* tým, že zhltne svet, aby zistil empirickú pravú podstatu tej priesvitnej tečúcej látky, a potom *vráti* nový pevne daný pojem Voda_{42} alebo H20, potom tento pojem pohlcujúci svet je ten istý na naše Zemi aj na Dvojčati Zeme; akurát na rôznych miestach dáva rôzne odpovede.

Ak si myslíte, že „voda" znamená H20, potom tento pojem nerobí nič iné, keď ho presúvame medzi svetmi, takže Dvojča Zeme neobsahuje žiadne H20.

A samozrejme nemá zmysel hádať sa, čo zvuk slabík „vo-da" naozaj znamená.

Mali by ste si teda vybrať jednu definíciu a používať ju konzistentne? Ale nie je také ľahké uchrániť sa pred zmätkom. Musíte sa naučiť *uvedomovať si* rozdiely medzi curryovanými a necurryovanými formami pojmov.

Keď vezmete necurryovaný pojem vody a aplikujete ho v inom svete, je to ten istý pojem, ale *odkazuje* na inú vec; to jest, aplikujeme konštantnú funkciu pohlcujúcu svety na iný svet, a dostávame inú návratovú hodnotu. Na Dvojčati Zeme „voda" je XYZ a nie je H20; na našej Zemi „voda" je H20 a nie je XYZ.

Na druhej strane, ak vezmete „voda" ako odkaz na to, čo by predchádzajúci mysliteľ nazval „výsledkom aplikovania pojmu ‚voda' na *našej* Zemi", potom na Dvojčati Zeme XYZ nie je voda, a H20 je.

Celá mätúcosť následnej filozofickej debaty sa zakladali na sklone *inštinktívne* pojmy curryovať alebo ich *inštinktívne* odcurryovať.

Podobne si vyžaduje krok navyše, aby si Fred uvedomil, že iní činitelia, napríklad Príšera s Hmyzími Očami, si budú vyberať unášanie na zneuctenie na základe Príťažlivosť_PHO(Žena) a nie Príťažlivosť_Fred(Žena). Aby toto urobil, musí si Fred vedome znovupredstaviť Príťažlivosť ako funkciu s dvoma argumentmi. Všetko, čo Fredov mozog robí inštinktom, je vyhodnocovanie Žena.príťažlivosť – to jest, Príťažlivosť_Fred(Žena); ale je to označené jednoducho Žena.príťažlivosť.

Pevná matematická funkcia Príťažlivosť_20934 nikde neodkazuje ani na Freda ani na PHO, iba na ženu, takže Fred *inštinktívne* nevidí, prečo by PHO mala vyhodnocovať „príťažlivosť" nejako inak. A naozaj, PHO by *nevyhodnocovala* Príťažlivosť_20934 nijako inak, keby sa z nejakého čudného dôvodu zaujímala o výsledok tejto konkrétnej funkcie; ale je *empirický* fakt o PHO, že *na rozhodovanie, koho uniesť*, používa inú funkciu.

Ak sa čudujete, čo je pointou tejto analýzy, budeme ju potrebovať neskôr, aby sme mohli hrať Tabu s takými mätúcimi slovami, ako je „objektívny", „subjektívny" a „ľubovoľný".

* →

270. Čo by ste robili bez morálky?

Tým, ktorí hovoria: „Nič nie je skutočné," som raz odpovedal: „To je skvelé, ale toto nič funguje?"

Predstavte si, že by ste sa dozvedeli, náhle a definitívne, že nič nie je morálne a nič nie je správne; že všetko je povolené a nič nie je zakázané.

Zdrujúca správa, to iste – a nie, ja vám nehovorím, že v skutočnom živote je to takto. Ale predstavte si, že by som vám to *povedal*. Predstavte si, že nech je základom vašej morálnej filozofie čokoľvek, že by som to presvedčivo roztrhal na kúsky, a navyše vám ukázal, že nič iné nemôže zaplniť jeho miesto. Predstavte si, že by som vám *dokázal*, že úžitok všetkého sa rovná nule.

Viem, že Vaša Morálna Filozofia je rovnako pravdivá a nevyvrátiteľná ako že 2 + 2 = 4. Ale aj tak vás žiadam, aby ste sa čo najlepšie pokúsili o tejto myšlienkový experiment, a konkrétne si predstavili tieto možnosti, aj keď vyzerajú bolestivé, alebo zbytočné, alebo logicky neumožňujúce žiadnu dobrú odpoveď.

→ http://lesswrong.com/lw/ro/2place_and_1place_words/

Stále by ste dávali sprepitné taxikárom? Podvádzali by ste vášho partnera alebo partnerku? Keby na vlakových koľajniciach ležalo dieťa v bezvedomí, stále by ste ho odtiaľ odtiahli?

Stále by ste jedli to isté jedlo ako doteraz – alebo by ste jedli iba to najlacnejšie jedlo, pretože nemá *zmysel* mať potešenie – alebo by ste jedli veľmi drahé jedlá, pretože nemá *zmysel* šetriť si peniaze na zajtra?

Obliekali by ste sa do čierneho a písali pochmúrne básne a odsudzovali všetkých altruistov ako bláznov? Ale neexistuje dôvod, prečo by ste to *mali* robiť – je to len uložená myšlienka.

Zostali by ste v posteli, pretože by nebol dôvod vstať? A čo potom, keby ste konečne od hladu vstali a vpotácali sa do kuchyne – čo by ste robili potom ako by ste dojedli?

Išli by ste si zase čítať *Overcoming Bias*, a ak nie, čo by ste čítali namiesto toho? Snažili by ste sa stále byť rozumní, a ak nie, čo by ste si namiesto toho mysleli?

Zavrite oči a doprajte si toľko času, koľko len treba na odpoveď:

Čo *by* ste robili, keby nič nebolo správne?

*→

271. Zmeniť svoju metaetiku

Ak poviete: „Zabíjať ľudí je zlé“, to je morálka. Ak poviete: „Nemali by ste zabíjať ľudí, lebo to Boh zakázal“ alebo „Nemali by ste zabíjať ľudí, lebo je to proti vesmírnym trendom“, to je metaetika.

Rovnako ako je viac zhody ohľadom špeciálnej relativity než ohľadom otázky: „Čo je to veda?“, je pre ľudí ľahšie zhodnúť sa na tom, že „Vražda je zlá“, než zhodnúť sa na tom, *čo* ju robí zlou, alebo čo to *znamená*, že niečo je zlé.

Ľudia sa často pripútajú k svojej metaetike. Naozaj často trvajú na tom, že ak je ich metaetika nesprávna, všetky morálka sa nevyhnutne rozpadne. Môže byť zaujímavé sledovať diskusiu metaetikov – veriacich, Objektivistov, Platonistov, atď. - ktorí sa všetci zhodnú, že zabíjať je zlé; ktorí sa všetci nezhodnú na tom, čo to znamená, že je niečo „zlé“; a ktorí trvajú na tom, že ak je ich metaetika nepravdivá, potom sa morálka rozpadne.

Je jasné, že mnohí ľudia, ak majú urobiť filozofický pokrok, budú musieť niekedy vo svojom živote zmeniť metaetiku. Možno to budete musieť urobiť *vy*.

V tomto bode môže byť užitočné mať otvorenú ústupovú líniu – nie ústup od morálky, ale ústup od Vašej Terajšej Metaetiky. (Veď viete, tej jedinej, ktorá, ak nie je pravdivá, nenecháva žiaden možný dôvod na nezabíjanie ľudí.)

A tak som nastavoval tieto únikové línie v mnohých rôznych článkov, zosumarizovaných nižšie. Pretože som sa naučil, že zmeniť metaetické názory je takmer nemožné v prítomnosti nezodpovedaného pripútania.

Ak napríklad niekto verí, že autorita „Nezabiješ“ pochádza od Boha, potom je niekoľko dobre známych vecí, ktoré im môžete povedať, a ktoré možno pomôžu nastaviť líniu ústupu – na rozdiel od okamžitého útoku na uveriteľnosť Boha. Môžete povedať: „Prevezmi osobnú zodpovednosť! Dokonca aj keby si dostal príkazy od Boha, bolo by to tvoje vlastné rozhodnutie poslúchať tieto príkazy. Dokonca aj keby si od Boha nedostal príkaz byť morálny, aj tak by si mohol jednoducho byť morálny.“

Horeuvedený argument v skutočnosti platí všeobecne pre dosť veľa metaetík – iba nahradíte Ich Obľúbený Zdroj Morálky alebo dokonca slovo „morálka“ namiesto „Boh“. Aj keby váš konkrétny zdroj morálnej autority zlyhal, nemohli by ste *aj tak* skrátka odtiahnuť to dieťa z koľajníc? A vôbec, kto je to, ak nie vy, kto sa vôbec najprv rozhodol nasledovať tento zdroj morálnej autority? Akú zodpovednosť tu naozaj odovzdávate?

Takže tá najdôležitejšia línia ústupu je: Ak vám vaša metaetika prestane hovoriť, že máte zachraňovať životy, aj tak môžete jednoducho odtiahnuť to dieťa z koľajníc. Aby som parafrázoval Piersa Anthonyho, iba tí, ktorí majú morálku, sa boja, či ju majú alebo nie. Keby vám vaša metaetika povedala,

→ http://lesswrong.com/lw/rq/what_would_you_do_without_morality/

aby ste zabíjali ľudí, prečo by ste ju vôbec *mali* počúvať? Možno to, čo by ste robili, aj keby neexistovala žiadna morálka, *je* vaša morálka.

Pointa samozrejme nie je, že žiadna morálka neexistuje; ale že môžete udržať svoju vôľu na mieste a nebáť sa, že stratíte zo zreteľa to, čo je pre vás dôležité, zatiaľ čo sa menia vaše predstavy o *podstate* morálky.

Iné články sú tu, aby nastavili línie ústupu konkrétne pre *naturalistickejšie* metaetiky. Radosť z púhej skutočnosti a Vysvetlenie verzus vyvrátenie tvrdia, že by ste nemali byť sklamaní žiadnou stránkou života len preto, že sa ukáže ako *vysvetliteľná* namiesto vnútorne tajomnej: lebo ak sa nedokážeme tešiť z iba skutočných vecí, naše životy budú iste prázdne.

Neexistujú univerzálne presvedčivé argumenty vytvára líniu ústupu od túžby, aby *každý* súhlasil s našimi morálnymi argumentmi. Existuje silná morálna intuícia, ktorá hovorí, že ak sú naše morálne argumenty správne, tak by sme ich predsa mali byť schopní ľuďom *vysvetliť*. To môže byť pravda medzi ľuďmi, ale nemôžete morálne argumenty vysvetľovať kameňu. Neexistuje ideálny študent filozofie dokonalej prázdnoty, ktorého by bolo možné presvedčiť, aby implementoval modus ponens, ak začal bez neho. Ak myseľ neobsahuje to, čím by sa dalo pohnúť vašimi morálnymi argumentmi, nebude na ne reagovať.

Ale nie je potom celá morálka kruhová logika, ktorá sa tým pádom rozpadne? Kde rekurzívne zdôvodňovanie narazí na dno a Môj druh reflexie vysvetľujú rozdiel medzi slučkou vnútorne konzistentnou na meta-úrovni a skutočnou kruhovou logikou. Nemali by ste zistiť, že si hovoríte: „Vesmír je jednoduchý, lebo je jednoduchý“ alebo „Vražda je zlá, lebo je zlá“; ale nemali by ste sa ani pokúšať zbaviť Occamovej britvy kým vyhodnocujete pravdepodobnosť, že Occamova britva funguje, ani by ste sa nemali pokúšať vyhodnotiť „Je vražda zlá?“ z niečoho mimo vášho mozgu. Neexistuje žiaden ideálny študent filozofie dokonalej prázdnoty, na ktorého by ste sa mohli odmotať – pokúste sa nájsť dokonalý kameň, na ktorom môžete stáť, a skončíte sám ako kameň. Takže namiesto toho používajte plnú silu svojej inteligencie, svoju plnú rozumnosť a plnú morálku, keď skúmate svoje vlastné základy.

Môžeme vytvoriť aj ústupovú líniu pre tých, ktorí sa boja dovoliť evolúcii, aby hrala *kauzálnu* rolu v ich predstave o tom, ako vznikla morálka. (Podotýkam, že toto je extrémne iné ako dať evolúcii postavenie *zdôvodnenia* v morálnych teóriách.) Láska musela nejako prísť do existencie – lebo ak sa nedokážeme tešiť z vecí, ktoré dokážu prísť do existencie, naše životy budú veru prázdne. Evolúcia nemusí byť príliš *príjemný* spôsob, ako sa láska vyvinie, ale hodnoťme výsledný produkt – nie zdroj. Inak sa dopustíte toho, čo sa označuje ako Klam pôvodu: príčina nie je to isté ako zdôvodnenie. Nemôžete predsa vystúpiť mimo mozgu, ktorý vám dala evolúcia: Rebelovanie proti prírode je možné iba v rámci prírody.

Dávnejšia séria o evolučnej psychológii by mala oslobodiť od metaetického mätúceho názoru, že každý normálny človek myslí na svoju reproduktívnu spôsobilosť, hoci len nevedome, keď sa rozhoduje. Iba evoluční biológovia vôbec vedia ako *definovať* genetickú spôsobilosť, a poznajú ju príliš dobre na to, aby si mysleli, že definuje morálku.

Znepokojujúca je aj samotná predstava, že morálka sa možno počíta vo vnútri našich vlastných myslí – nevyplýva z toho azda, že morálka je púha myšlienka? Nevyplýva z toho, že čokoľvek si myslíte, že je správne, musí byť správne?

Nie. Len preto, že sa nejaká hodnota počíta vnútri vašej hlavy, nemusí to znamenať, že vypočítaná hodnota je o vašich myšlienkach. Je rozdiel medzi kalkulačkou, ktorá počíta „Koľko je 2 + 3“ a „Čo vypíšem, keď niekto stlačí ‚2‘, ‚+‘ a ‚3‘?“

A nakoniec, ak vám život pripadá bolestivý, redukcionizmus nemusí byť skutočnýmm zdrojom vášho problému – ak vám život vo svete skladajúcom sa iba z častíc pripadá príliš neznesiteľný, možno samotný váš život nie je dostatočne vzrušujúci?

A ak sa čudujete, prečo považujem túto záležitosť s metaetikou za dôležitú, keď to nakoniec aj tak všetko dá morálnu normálnosť... povie vám, že máte odtiahnuť dieťa z koľají, a nie naopak...

Nuž, *existuje* opozícia voči rozumnosti, od ľudí, ktorí si myslia, že z vesmíru odčerpáva zmysel.

A toto je špeciálny prípad všeobecného javu, kde sa mnoho mnoho ľudí prevráti hore nohami vďaka neporozumeniu, odkiaľ pochádza ich morálka. Slabá metaetika tvorí časť učenia nejednej sekty, vrátane tých veľkých. Moja cieľová skupina nie sú iba ľudia, ktorí sa boja, že život je nezmyselný, ale aj tí, ktorí došli k záveru, že láska je ilúzia, pretože skutočná morálka musí zahŕňať maximalizáciu vašej inkluzívnej spôsobilosti, alebo tí, ktorí došli k záveru, že neopätovaná láskavosť je zlá, pretože skutočná morálka vychádza iba zo sebectva, atď.

*→

272. Mohlo by hocičo byť správne?

Pred rokmi bol Eliezer_{1999} presvedčený, že nevie nič o morálke.

Pokiaľ vedel, morálka si mohla vyžadovať vyhubenie ľudského druhu; a keby to tak bolo, nevidel žiadnu cnosť vo vzdorovaní morálke, pretože si myslel, že podľa definície, ak predpokladá takýto morálny fakt, by to znamenalo, že vyhubenie ľudstva je to, čo „by sa malo“ urobiť.

Myslel som si, že by som azda vedel *zistiť*, čo je správne, keby som mal dosť času na rozmýšľanie a dosť faktov, ale že o tom teraz nemám žiadne informácie. Nemohol som dôverovať evolúcii, ktorá ma zostavila. Aký základ mi to nechávalo, na ktorom by som mohol stáť?

Nuž, Eliezer_{1999} sa ohromne mýlil ohľadom podstaty morálky, pokiaľ išlo o jeho explicitne reprezentovanú filozofiu.

Ale ako raz poznamenal Davidson, ak veríte, že „bobry“ žijú v púšti, majú čisto bielu farbu, a vážia v dospelosti 150 kíl, potom nemáte žiadne názory *o* bobroch, pravdivé ani nepravdivé. Aspoň niektoré z vašich názorov musia byť správne, aby tie zvyšné mohli byť nesprávne *o* niečom.[244]

Môj názor, že nemám *žiadne* informácie *o* morálke, bol vnútorne nekonzistentný.

Hovoriť, že nič neviem, mi pripadalo cnostné, pretože ma raz naučili, že je cnostné vyznávať svoju nevedomosť. „Viem iba to, že nič neviem“ a tak ďalej. Ale v tomto prípade by som bol na tom lepšie, keby som zvážil nesporne prehnané porekadlo: „Najväčším bláznom je ten, kto si neuvedomuje, že je múdry.“ (Toto sa ani len nepribližuje k *najväčšiemu* druhu bláznovstva, ale je to istý druh bláznovstva.)

Je zlé zabíjať ľudí? No, myslel som si to, ale nebol som si istý; možno bolo správne zabíjať ľudí, hoci to vyzeralo menej pravdepodobne.

Aký druh *procedúry* by mohol odpovedať, či je správne zabíjať ľudí? Ani to som nevedel; ale myslel som si, že keby sme postavili všeobecnú superinteligenciu (to, čo by som neskôr označil ako „duch dokonalej prázdnoty“), tá by vedela, chápete, uvažovať o tom, čo je pravdepodobne správne a nesprávne; a keďže by bola *superinteligentná*, musela byť dôjsť na správnu odpoveď.

Problém, na ktorý sa mi akosi darilo príliš silno nemyslieť, bol, kde by táto superinteligencia získala procedúru, ktorá by objavila procedúru, ktorá by objavila procedúru, ktorá by objavila morálku – keby som to ja nemohol napísať do počiatočného stavu, ktorý napísal nasledujúcu UI, ktorá napísala nasledujúcu UI.

Ako neskôr povedal Marcello Herreshoff : „Unúvame sa spustiť počítačový program, iba keď nevieme výsledok, ale vieme dôležitý fakt o výsledku.“ Ak som nevedel nič o morálke, a dokonca som ani netvrdil, že poznám podstatu morálky, ako by som potom mohol zostaviť akýkoľvek počítačový program – hoci aj „superinteligentný“ alebo „sebazdokonaľujúci“ - a tvrdiť, že vypíše niečo, čo sa volá „morálka“?

V informatike existujú vety neexistuje-obed-zadarmo – vo vesmíre maximálnej entropie nie je žiaden plán v priemere lepší než hocijaký iný. Ak nemáte vôbec žiadnu vedomosť o „morálke“, potom neexistuje ani žiadna výpočtová procedúra, ktorá by mala väčšiu pravdepodobnosť než iné vypočítať „morálku“, a žiadna meta-procedúra, ktorá by mala väčšiu pravdepodobnosť než iné vypočítať procedúru, ktorá vypočíta „morálku“.

→ http://lesswrong.com/lw/sk/changing_your_metaethics/

244 Rorty, „Out of the Matrix: How the Late Philosopher Donald Davidson Showed That Reality Can’t Be an Illusion.“

Myslel som si, že aj duch dokonalej prázdnoty by po zistení, že nevie nič o morálke, iste videl ako morálny imperatív *rozmýšľať o morálke*.

Ťažkosť však spočíva v slove *myslieť*. Myslenie nie je aktivita, ktorú by duch dokonalej prázdnoty automaticky dokázal vykonávať. Myslenie si vyžaduje zbiehanie nejakého *konkrétneho* výpočtu, ktorý je myšlienkou. Aby sa reflexívna UI rozhodla myslieť, vyžaduje, aby poznala nejaký výpočet, o ktorom verí, že má *väčšiu* pravdepodobnosť povedať je to, čo chce vedieť, než konzultovanie s doskou Ouija; UI musí mať aj predstavu, ako interpretovať výsledok.

Ak niekto nevie nič o morálke, čo vôbec znamená slovo „malo by sa“? Ak neviete, či je smrť správna alebo nesprávna – a neviete, ako by sa dalo zistiť, či je smrť správna alebo nesprávna – a neviete, či ľubovoľná konkrétna procedúra môže *vypísať* procedúru, ktorá povie, či je smrť správna alebo nesprávna – čo potom tieto slová „správna“ a „nesprávna“ vôbec *znamenajú*?

Ak slová „správne“ a „nesprávne“ nemajú v sebe *nič* zabudované – žiaden počiatočný bod – ak je *všetko* ohľadom morálky voľne k dispozícii, nie iba obsah, ale aj štruktúra a počiatočný bod a rozhodovacia procedúra – aký to má potom význam? V čom je rozdiel medzi: „Neviem, čo je správne“ a „Neviem, čo je wakalixa“?

Vedec môže povedať, že vo vede je všetko voľne k dispozícii, pretože ľubovoľnú teóriu možno vyvrátiť; ale predsa má nejakú predstavu o tom, čo by sa počítalo ako *indícia*, ktorá by mohla teóriu vyvrátiť. Mohlo by existovať niečo, čo by zmenilo, čo vedec považuje za indíciu?

No vlastne áno; vedec, ktorý prečítal trochu Karla Poppera a myslel si, že vie, čo znamená „indícia“, by sa mohol dozvedieť o dôkazoch koherencie a jedinečnosti, na ktorých stojí bayesovská pravdepodobnosť, a to by mohlo zmeniť jeho definíciu indície. Možno predtým ani nemal žiadnu *explicitnú predstavu*, že by takýto dôkaz mohol existovať. Ale mal by implicitnú predstavu. Bolo by zabudované v jeho mozgu, ak aj nie explicitne reprezentované, že taký a taký argument by ho naozaj presvedčil, že bayesovská pravdepodobnosť dáva lepšiu definíciu „indície“ než tá, ktorú používal dovtedy.

Rovnako by ste mohli povedať: „Neviem, čo je morálka, ale rozoznám ju, keď ju uvidím,“ a dávalo by to zmysel.

Ale potom sa celkom nebúrite proti svojej vlastnej vyvinutej povahe. Predpokladáte, že čokoľvek, čo je vo vás zabudované, aby rozoznalo „morálku“ je, ak nie absolútne dôveryhodné, tak prinajmenšom váš počiatočný stav, v ktorom začnete debatovať. Môžete dôverovať svojim morálnym intuíciám, že vám dajú *vôbec nejakú* informáciu o morálke, ak sú produktom čírej evolúcie?

Lenže ak odhodíte každú procedúru, ktorú vám evolúcia dala, *aj všetky jej produkty*, potom odhodíte celý svoj mozog. Odhodíte všetko, čo by potenciálne mohlo rozoznať morálku, keď ju to uvidí. Odhodíte všetko, čo by potenciálne mohlo reagovať na morálne argumenty aktualizovaním vašej morálky. Odmotáte dokonca samotného odmotávača: odhodíte intuície, na ktorých sa zakladá váš záver, že *nemôžete dôverovať evolúcii*, že bude morálna. To vaše *terajšie* morálne intuície vám hovoria, že evolúcia nevyzerá ako veľmi *dobrý* zdroj morálky. Čo potom budú slová „správne“ a „malo by sa“ a „lepšie“ vôbec *znamenať*?

Ľudia nedokážu dokonale rozoznať pravdu, keď ju vidia, a lovci-zberači nemali explicitný pojem bayesovského kritéria indície. Ale celá naša veda a celá naša teória pravdepodobnosti boli postavené navrchu reťaze odvolávok na našu inštinktívnu predstavu „pravdy“. Keby toto jadro bolo skazené, neexistovalo by nič, čo by sme mohli *v princípe* urobiť, aby sme došli k terajšej predstave vedy; predstava vedy by skrátka znela celkom nepríťažlivo a nezmyselne.

Jeden z argumentov, ktoré by možno mohli prebrať moje puberťácke ja z jeho chyby, keby som sa mohol vrátiť v čase a porozprávať sa s ním, by bola otázka:

Môže existovať nejaká morálka, nejaké dané správne a nesprávne, ktorú ľudia nevnímajú, nechcú vnímať, nebudú vidieť žiaden presvedčivý morálny argument za jej prijatie, ani žiaden morálny argument za prijatie procedúry, ktorá ju prijme, a tak ďalej? Môže existovať nejaká morálka, a my *celkom* mimo je rámca referencie? Ale čo potom robí z tejto veci *morálku* – namiesto nejakej kamennej dosky niekde s vytesanými slovami „nezabiješ“, bez absolútne akéhokoľvek *zdôvodnenia*?

Takže toto všetko naznačuje, že by ste mali byť ochotní pripustiť, že možno viete *niečo* o morálke. Azda nič nespochybniteľné, ale počiatočný stav, s ktorým sa môžete začať pýtať sami seba. Azda zabudované vo vašom mozgu, ale vám explicitne nie známe; ale aj tak, to, čo *by* váš mozog rozoznal ako *správne*, je to, *o* čom hovoríte. Prinajmenšom ten spôsob, ktorým *reagujete na morálne argumenty*, vezmite ako *začiatočný bod*, aby ste identifikoval „morálku“ ako niečo, o čom rozmýšľate.

Ale to je pomerne veľký krok.

Vyplýva z neho, že by ste mali prijať svoju vlastnú myseľ ako morálny referenčný rámec, a nie že všetka morálka je veľké svetlo žiariace odniekiaľ mimo (ktoré by ste v princípe nemuseli byť vôbec schopní vnímať). Vyplýva z neho, že aj keby existovalo nejaké svetlo a váš mozog by sa ho rozhodol rozoznávať ako „morálku“, stále by to bol váš vlastný mozog, kto to rozoznáva, a vy by ste sa nevyhli kauzálnej zodpovednosti – ani morálnej zodpovednosti, z môjho pohľadu.

Vyplýva z neho, aby ste odhodili predstavu, že by duch dokonalej prázdnoty s vami nevyhnutne súhlasil, pretože tento duch sa môže nachádzať v inom morálnom referenčnom rámci, reagovať na iné argumenty, *klásť si iné otázky*, keď počíta čo-urobiť-ďalej.

A ak ste ochotní zabudovať aspoň pár vecí do samotného zmyslu tejto témy o „morálke“, táto vlastnosť *správnosti*, o ktorej hovoríte, keď hovoríte, že niečo je „správne“ - ak ste ochotní pripustiť dokonca aj, že morálka je to, o čom sa hádate, keď sa hádate o „morálke“ - prečo potom nepripustiť aj iné intuície, iné časti seba, do tohto štartového bodu?

Prečo nepripustiť, že, *ceteris paribus*, radosť je lepšia než smútok?

Neskôr možno nájdete nejaký základ v sebe alebo postavený na sebe, pomocou ktorého toto skritizujete – ale prečo to neprijať aspoň zatiaľ? Nie len ako osobnú preferenciu, pripomínam; ale ako niečo zabudované do *otázky*, ktorú sa pýtate, keď sa pýtate: „Čo je naozaj správne?“

Ale potom by ste mohli zistiť, že viete o morálke celkom veľa! Nič naisto – nič nespochybniteľné – nič nediskutovateľné – ale aj tak, celkom dosť informácií. Ste ochotní vzdať sa svojej sokratovskej nevedomosti?

Nejdem argumentovať z definície, samozrejme. Ale ak tvrdíte, že o morálke neviete vôbec nič, potom budete mať problémy s *významom* svojich slov, nielen s ich dôveryhodnosťou.

*→

273. Morálka ako pevne daný výpočet

Toby Ord napísal komentár:

> Eliezer, práve som si znovu prečítal tvoj článok a rozmýšľam, či je toto dobré stručné zhrnutie tvojho postoja (vynechajúc to, ako si sa k nemu dostal):
>
> „Mal by som X“ znamená, že by som sa pokúsil o X, keby som mal všetky informácie.

Toby je odborník, takže ak to nepochopil on, musím to vysvetliť znovu. Skúsim iný postup vysvetľovania – bližší k historickej ceste, ako som sa k svojmu postoju dostal.

Predstavte si, že postavíte UI a – ignorujúc, že systém cieľov UI nemôžete postaviť na anglických vetách, a že všetky takéto popisy sú iba sen – skúsite do UI vložiť princíp rozhodujúci o jej konaní: „Rob to, čo ja chcem.“

A predpokladajme, že ste dizajn UI *dosť* utrafili – že to skrátka neskončí tým, že vyplní vesmír kancelárskymi spinkami, tvarohovými tortami, ani drobnými molekulárnymi kópiami spokojných programátorov – a že jeho funkcia úžitku naozaj priraďuje úžitok stavom sveta spôsobom, ktorý by sme po anglicky opísali takto:

> <Programátor slabo chce „X“, existuje 20 X>: +20
>
> <Programátor silno chce „Y“, existuje 20 X>: 0

→ http://lesswrong.com/lw/sb/could_anything_be_right/

<Programátor slabo chce „X“, existuje 30 Y>: 0

<Programátor silno chce „Y“, existuje 30 Y>: +60

Samozrejme vidíte, že toto zničí celý svet.

...pretože ak programátor na začiatku slabo chcel „X“, ale X sa ťažko dosahuje, UI zmení samotného programátora, aby silno chcel „Y“, ktoré sa vyrába ľahko, a potom vyrobí hromadu Y. Referentom „Y“ môžu byť povedzme atómy železa – tie sú vysoko stabilné.

Dokážete zaplátať tento problém? Nie. Ako všeobecné pravidlo, chybné dizajny Priateľských UI sa nedajú zaplátať.

Ak sa pokúsite ohraničiť funkciu úžitku, alebo urobíte, že sa UI nebude starať o to, ako *veľmi* programátor niečo chce, UI má stále motív (ako maximalizátor *očakávaného* úžitku) urobiť, aby programátor chcel niečo, čo sa dá dosiahnuť s veľmi vysokým stupňom istoty.

Ak sa to pokúsite urobiť tak, že UI nemôže zmeniť programátora, potom sa UI nemôže s programátorom rozprávať (lebo rozprávanie sa s niekým ho zmení).

Ak sa pokúsite zakázať konkrétnu množinu spôsobov, ako by UI mohla zmeniť programátora, UI má motív superinteligentne hľadať medzery a spôsoby ako zmeniť programátora nepriamo.

Ako všeobecné pravidlo, chybné dizajny PUI sa nedajú zaplátať.

My sami si nepredstavujeme budúcnosť a nesúdime, že ľubovoľná budúcnosť, v ktorej naše mozgy niečo chcú a táto vec existuje, je dobrá budúcnosť. Keby sme rozmýšľali takto, povedali by sme: „Hurá! Poď nás zmeniť tak, aby sme silno chceli niečo lacné!“ Lenže my toto *nehovoríme*, čo znamená, že tento dizajn UI je *od základu* chybný: bude vyberať veci, ktoré by sme si s najväčšou pravdepodobnosťou nevybrali; bude hodnotiť žiadúcnosť veľmi odlišne od toho, ako ju hodnotíme my. Túto disharmóniu v jadre nemožno zaplátať tým, že zakážeme niekoľko konkrétnych spôsobov zlyhania.

Existuje aj dualita medzi problémami Priateľskej UI a problémami morálnej filozofie – musíte však túto dualitu postaviť celkom presným spôsobom. Ak teda chcete, jadrom problému je to, že si UI bude vyberať spôsobom, ktorý sa veľmi líši od štruktúry toho, čo je, však viete, naozaj *správne* – nehľadiac na to, ako si vyberáme my. Nie je celou pointou tohto problému, že niečo iba *chcieť* to ešte *nerobí* správnym?

Takže toto je tá paradoxne vyzerajúca vec, ktorú som prirovnal k rozdielu medzi:

Kalkulačkou, ktorá keď stlačíte „2“, „+“ a „3“, skúša vypočítať:

„Koľko je 2 + 3?“

A kalkulačkou, ktorá keď stlačíte „2“, „+“ a „3“, skúša vypočítať:

„Čo vypíše táto kalkulačka, keď stlačíte ‚2‘, ‚+‘ a ‚3‘?“

Kalkulačka typu 1 by *chcela* vypísať 5.

„Kalkulačka“ typu 2 by mohla vrátiť ľubovoľný výsledok; a samotným činom vrátenia tohto výsledku sa to *stane* správnou odpoveďou na otázku, ktorú si vo vnútri položila.

My sme ako tá kalkulačka typu 1. Ale hypotetickú UI staviame, ako keby mala odrážať kalkulačku typu 2.

Teraz si predstavte, že sa kalkulačka typu 1 pokúša postaviť UI, akurát že kalkulačka typu 1 *nepozná* odpoveď na svoju vlastnú otázku. Tá kalkulačka si túto otázku kladie zo svojej podstaty, narodila sa, aby si kládla túto otázku, bola už stvorená v pohybe s touto otázkou – ale táto kalkulačka nerozumie svojim vlastným tranzistorom; nedokáže vypísať odpoveď, ktorá je veľmi komplikovaná a nemá žiadnu jednoduchú aproximáciu.

Takže táto kalkulačka chce postaviť UI (je to pomerne bystrá kalkulačka, akurát nemá prístup ku svojom vlastným tranzistorom) a chce, aby UI dala správnu odpoveď. Táto kalkulačka však nedokáže vypísať túto otázku. Preto kalkulačka chce, aby sa UI pozrela na ňu, kde je napísaná táto otázka, a dala jej odpoveď, ktorú nájde implicitne v týchto tranzistoroch. To sa však nedá urobiť pomocou lacnej skratky

vo funkcii úžitku, ktorá povie: „Pre každé X: <kalkulačka sa pýta ‚X?‘, odpovedz X>: úžitok 1; inak: úžitok 0“, pretože toto v skutočnosti odráža funkciu úžitku kalkulačky typu 2, nie kalkulačky typu 1.

Toto nás dostane k otázkam PUI, do ktorých nebudem zachádzať (na niektorých z nich sám stále pracujem).

Keď však odstúpime od detailov dizajnu PUI a vrátime sa späť do pohľadu morálnej filozofie, potom *sme sa doteraz rozprávali* o duále k morálnej otázke: „Ak je ‚správne‘ iba preferencia, potom čokoľvek, čo niekto chce, je ‚správne‘.“

Kľúčovým pojmom je myšlienka, že čo označujeme slovom „správne“ je *pevne daná* otázka, alebo skôr *pevne daný rámec*. Môžeme natrafiť na morálne argumenty, ktoré zmenia naše konečné hodnoty, a dokonca aj na morálne argumenty, ktoré zmenia, čo považujeme za morálny argument; ale aj tak to všetko vyrastá z konkrétneho štartového bodu. Nevnímame to tak, že máme v sebe otázku: „Čo sa rozhodnem urobiť?“, čo by bola kalkulačka typu 2; čokoľvek, pre čo by sme sa rozhodli, by sa tým stalo správne. Vnímame to tak, že máme v sebe otázku: „Čo zachráni mojich priateľov a mojich blízkych pred zranením? Ako môžeme byť všetci ešte šťastnejší? ...“, kde to „...“ je asi tisíc ďalších vecí.

Takže „mal by som X“ neznamená, že by som urobil X, keby som mal všetky informácie.

„Mal by som X“ znamená, že X je odpoveďou na otázku: „Čo zachráni mojich blízkych? Ako môžeme byť všetci ešte šťastnejší? Ako získať viac kontroly nad svojím životom? Aké najzábavnejšie vtipy môžeme povedať? ...“

A možno ani *neviem*, čo v skutočnosti táto otázka *je*; možno si nedokážem vytlačiť svoj terajší odhad ani rámec okolo mňa; ale viem, ako všetci, čo nie sú morálni relativisti, inštinktívne vedia, že tá otázka určite nie je iba: „Ako môžem urobiť hocičo, čo chcem?“

Keď vám tieto dve formulácie začnú pripadať rovnako odlišné ako „sneh“ a sneh, potom budete mať vytvorené rôzne vedrá pre citáciu a referent.

*→

274. Čarovné kategórie

> Môžeme navrhnúť inteligentné stroje tak, že ich primárna, vrodená emócia bude nepodmienená láska voči všetkým ľuďom. Najprv postavíme pomerne jednoduché stroje, ktoré sa naučia rozoznávať šťastie a nešťastie vo výrazoch ľudskej tváre, v ľudskom hlase, a ľudskej reči tela. Potom môžeme napevno zabudovať výsledok tohto učenia ako vrodenú emocionálnu hodnotu zložitejších inteligentných strojov, ktoré budú pozitívne podmieňované, keď budeme šťastní, a negatívne podmieňované, keď budeme nešťastní.
>
> --Bill Hibbard (2001), Superinteligentné stoje[245]

Toto bolo uverejnené v recenzovanom odbornom časopise, a autor o tom neskôr napísal celú knihu, takže to nie je slamený panák, o čom tu diskutujem...

Takže.... hm... čo by sa asi tak mohlo pokaziť...

Keď som spomenul (časť 7.2)[246], že Hibbardova UI by skončila vydláždením celej galaxie drobnými molekulárnymi usmievavými tvárami, Hibbard napísal rozhorčenú odpoveď so slovami:

> Keď bude reálne postaviť superinteligenciu, bude reálne postaviť pevne zabudované rozoznávanie „výrazov ľudskej tváre, ľudského hlasu, a ľudskej reči tela“ (aby som použil moje slová, ktoré citujete), ktoré presiahne presnosť rozoznávania dnešných ľudí ako sme vy a ja, takže určite nebude oblbnuté „drobnými molekulárnymi usmievavými tvárami“. Nemali

→ http://lesswrong.com/lw/sw/morality_as_fixed_computation/

245 Bill Hibbard, „Super-Intelligent Machines,“ *ACM SIGGRAPH Computer Graphics* 35, no. 1 (2001): 13–15, http://www.siggraph.org/publications/newsletter/issues/v35/v35n1.pdf.

246 Eliezer Yudkowsky, „Artificial Intelligence as a Positive and Negative Factor in Global Risk,“ in Bostrom and Ćirković, *Global Catastrophic Risks*, 308–345.

> by ste predpokladať takú biednu implementácia môjho nápadu, že nedokáže rozoznať veci, ktoré sú pre dnešných ľudí triviálne.

Keďže Hibbard ďalej napísal: „Takéto očividne protichodné predpoklady ukazujú, že Yudkowsky dáva prednosť dráme pred rozumom," ja tiež na rovinu poviem, že Hibbard ilustruje kľúčový bod: Neexistuje žiaden profesionálny certifikačný test, ktorým musíte prejsť predtým ako vám bude dovolené hovoriť o morálke UI. Ale to nie je moja dnešná prvoradá téma. Aj keď je kritickým faktom o dnešnom stave hry, že väčšina rádoby odborníkov na Všeobecnú UI / Priateľskú UI je na túto úlohu tak *naprosto* nepripravených, že nepoznám nikoho dostatočne cynického, aby si túto hrôzu predstavil skôr než uvidí na vlastné oči. Dokonca aj Michael Vassar bol asi po prvýkrát prekvapený.

Nie, dnes tu ie som, aby som rozoberal vetu: „Nemali by ste predpokladať takú biednu implementáciu môjho nápadu, že nedokáže rozoznať veci, ktoré sú pre dnešných ľudí triviálne."

Kde bolo, tam bolo – počul som tento príbeh v niekoľkých verziách a na niekoľkých miestach, občas citovaný ako fakt, ale nikdy som nevystopoval pôvodný zdroj – kde bolo, tam bolo, hovorím, chcela americká armáda použiť neurónové siete na automatické odhaľovanie zamaskovaných nepriateľských tankov.

Výskumníci vycvičili neurónovú sieť na 50 fotografiách zamaskovaných tankov medzi stromami, a 50 fotografiách stromov bez tankov. Pomocou štandardných techník učenia so supervíziou výskumníci vycvičili túto neurónovú sieť do vyváženia, ktoré správne ohodnocovalo cvičnú sadu – dalo výstup „áno" pre 50 fotografií zamaskovaných tankov, a výstup „nie" pre 50 fotografií lesa.

Toto však nedokazovalo, ani len nenaznačovalo, že nové príklady budú zatriedené správne. Neurónová sieť sa mohla aj „naučiť" 100 špeciálnych prípadov, ktoré by sa nezovšeobecnili na nové problémy. Nie „maskované tanky verzus les" ale iba „fotografia číslo 1 áno, fotografia číslo 2 nie, fotografia číslo 3 nie, fotografia číslo 4 áno..."

Výskumníci však múdro pôvodne urobili 200 fotografií, 100 fotografií tankov a 100 fotografií stromov, a v trénovacej sade použili iba prvú polovicu. Výskumníci spustili neurónovú sieť na zvyšných 100 fotografiách a neurónová sieť *bez ďalšieho tréningu* zatriedila všetky zostávajúce fotografie správne. Úspech potvrdený!

Výskumníci odovzdali dokončené dielo do Pentagonu, odkiaľ im ho čoskoro vrátili so sťažnosťou, že v ich vlastných testoch sa neurónovej sieti v rozlišovaní fotografií nedarí o nič lepšie než náhodnému výberu.

Ukázalo sa, že v dátovej množine výskumníkov boli fotografie zamaskovaných tankov urobené v zamračený deň, kým fotografie prázdneho lesa boli urobené v slnečný deň. Neurónová sieť sa naučila rozoznávať zamračený deň od slnenčného, namiesto rozoznávania zamaskovaného tanku od prázdneho lesa.

Toto podobenstvo – ktoré môže ale nemusí byť skutočné – znázorňuje jeden z najzákladnejších problémov v oblasti učenia so supervíziou, a vlastne v celej oblasti umelej inteligencie: Ak je medzi tréningovými a skutočnými problémami čo len najmenší rozdiel v kontexte – ak nie sú vylosované z toho istého nezávislého procesu s rovnakým rozdelením – minulé úspechy nie sú štatistickou zárukou budúcich úspechov. Nezáleží na tom, či sa zdá, že UI funguje dobre v cvičných podmienkach. (Toto nie je *neriešiteľný* problém, ale je to problém *neriešiteľný záplatou*. Existujú hlbšie spôsoby, ako tomu čeliť – čo je téma mimo rozsahu tohto článku – ale nie náplaste.)

Ako je opísané v kapitole Superexponenciálny priestor pojmov, existuje exponenciálne viac možných pojmov než je možných predmetov, rovnako ako je počet možných predmetov exponenciálny voči počtu vlastností. Ak má strana čiernobieleho obrázku 256 pixelov, potom celý obrázok má 65536 pixelov. Počet možných obrázkov je 2^{65536}. A počet možných *pojmov*, ktoré triedia obrázky na pozitívne a negatívne prípady – počet možných *hraníc*, ktoré by ste mohli nakresliť v priestore obrázkov – je 2^{265536}. Z tohto vidíme, že ešte aj učenie so supervíziou je takmer výhradne otázkou induktívneho sklonu, bez ktorého by bolo treba aspoň 2^{65536} zatriedených príkladov, aby sme rozlíšili medzi 2^{265536} možnými pojmami – dokonca aj keby klasifikácia zostávala v čase rovnaká.

Vráťme sa teda opäť k:

> Najprv postavíme pomerne jednoduché stroje, ktoré sa naučia rozoznávať šťastie a nešťastie vo výrazoch ľudskej tváre, v ľudskom hlase, a ľudskej reči tela. Potom môžeme napevno zabudovať výsledok tohto učenia ako vrodenú emocionálnu hodnotu zložitejších inteligentných strojov, ktoré budú pozitívne podmieňované, keď budeme šťastní, a negatívne podmieňované, keď budeme nešťastní.

a

> Keď bude reálne postaviť superinteligenciu, bude reálne postaviť pevne zabudované rozoznávanie „výrazov ľudskej tváre, ľudského hlasu, a ľudskej reči tela“ (aby som použil moje slová, ktoré citujete), ktoré presiahne presnosť rozoznávania dnešných ľudí ako sme vy a ja, takže určite nebude oblbnuté „drobnými molekulárnymi usmievavými tvárami“. Nemali by ste predpokladať takú biednu implementácia môjho nápadu, že nedokáže rozoznať veci, ktoré sú pre dnešných ľudí triviálne.

Rozlíšiť medzi nejakou fotografiou zamaskovaného tanku a nejakou fotografiou prázdneho lesa je triviálne v tom zmysle, že je možné povedať, že tieto dve fotografie nie sú totožné. Sú to rôzne polia pixelov s rôznymi nulami a jednotkami. Rozlíšiť medzi nimi je rovnako jednoduché ako skontrolovať, či sú tieto polia rovnaké.

Klasifikovať nové fotografie do pozitívnych a negatívnych príkladov „úsmevu“ pomocou usudzovania podľa množiny tréningových fotografií klasifikovaných ako pozitívne alebo negatívne, je celkom iný rád problému.

Keď máte obrázok 256 × 256 z ozajstného fotoaparátu, a ukáže sa, že tento obrázok znázorňuje zamaskovaný tank, nie je tam *dodatočný* 65537-y bit označujúci pozitívnosť – žiadna drobná značka XML, ktorá hovorí: „Tento obrázok je zo svojej podstaty pozitívny“. Je to iba pozitívny príklad z nejakého *konkrétneho* pojmu.

Ale pre hocijaké nie-Obrovské množstvo tréningových údajov – pre hocijaké tréningové údaje, ktoré neobsahujú na bit *presne* rovnaký obrázok, aký práve vidíte – existuje *superexponenciálne* veľa možných pojmov, ktoré sú v súlade s predchádzajúcimi klasifikáciami.

Pre UI je vyberanie alebo váženie spomedzi superxponenciálnych možností otázkou induktívneho skreslenia. Ktoré nemusí zodpovedať tomu, na čo myslel používateľ. Hranica medzi dvoma procesmi na klasifikovanie obrázkov – na jednej strane indukcia, na druhej strane používateľov skutočný cieľ – nemusí byť triviálne prekročiteľná.

Povedzme, že tréningové údaje UI sú:

> Databáza 1:
>
> \+ Usmev_1, Usmev_2, Usmev_3
>
> \- Zamracenie_1, Macka_1, Zamracenie_2, Zamracenie_3, Macka_2, Lod_1, Auto_1, Zamracenie_5

Teraz táto UI vyrastie na superinteligenciu a natrafí na tieto údaje:

> Databáza 2:
>
> Zamracenie_6, Macka_3, Usmev_4, Galaxia_1, Zamracenie_7, Nanotovaren_1, Molekularny_smajlik_1, Macka_4, Molekularny_smajlik_2, Galaxia_2, Nanotovaren_2

Nie je vlastnosťou *týchto databáz*, že vydedukovaná databáza, ktorú *by ste vy chceli*, je:

> \+ Usmev_1, Usmev_2, Usmev_3, Usmev_4
>
> \- Zamracenie_1, Macka_1, Zamracenie_2, Zamracenie_3, Macka_2, Lod_1, Auto_1, Zamracenie_5, Zamracenie_6, Macka_3, Galaxia_1, Zamracenie_7, Nanotovaren_1, Molekularny_smajlik_1, Macka_4, Molekularny_smajlik_2, Galaxia_2, Nanotovaren_2

a nie povedzme

+ Usmev_1, Usmev_2, Usmev_3, Molekularny_smajlik_1, Molekularny_smajlik_2, Usmev_4

- Zamracenie_1, Macka_1, Zamracenie_2, Zamracenie_3, Macka_2, Lod_1, Auto_1, Zamracenie_5, Zamracenie_6, Macka_3, Galaxia_1, Zamracenie_7, Nanotovaren_1, Macka_4, Galaxia_2, Nanotovaren_2

Obe tieto klasifikácie sú v súlade s tréningovými údajmi. Počet *pojmov* kompatibilných s tréningovými údajmi bude omnoho väčší, keďže viac než jeden pojem dokáže do spojenej databázy premietnuť rovnaký tieň. Ak priestor možných pojmov zahŕňa priestor možných výpočtov, ktoré klasifikujú príklady, tento priestor je nekonečný.

Ktorú klasifikáciu si UI vyberie? To nie je vnútornou vlastnosťou tréningových údajov; je to vlastnosť toho, ako UI urobí indukciu.

Ktorá je *správna* klasifikácia? To nie je vlastnosť tréningových údajov; je to vlastnosť vašich preferencií (alebo, ak chcete, vlastnosť idealizovanej abstraktnej dynamiky, ktorú nazývate „správne“).

Pojem, ktorý *ste chceli*, vrhá svoj tieň na tréningové údaje, ako ste vy sami označili každý prípad + alebo -, použitím svojej vlastnej inteligencie a preferencií. To je to, o čom je celé učenie so supervíziou – poskytovanie UI označených tréningových príkladov, ktoré obsahujú tiež kauzálneho procesu, ktorý vytvoril tieto označenia.

Ale pokiaľ tréningové údaje nevyberáte z *presne* rovnakého kontextu ako je skutočný život, tieto tréningové údaje budú v istom zmysle „plytké“, premietanie z omnoho viac-rozmerného priestoru možností.

Táto UI počas svojej tréningovej fázy, kým bola hlúpejšia než človek, nikdy nevidela drobnú molekulárnu usmievavú tvár, ani nevidela drobného činiteľa s počítadlom šťastia nastaveným na 10^{10100}. Áno, *vy*, keby ste nakoniec dostali drobný molekulárny smajlík – alebo povedzme veľmi realistickú drobnú sochu ľudskej tváre – hneď by ste vedeli, že toto nie je to, čo *vy* chcete počítať ako úsmev. Ale tento úsudok odráže neprirodzenú kategóriu, takú, ktorej klasifikujúca hranica citlivo závisí od vašich zložitých hodnôt. Sú to vaše vlastné plány a túžby, ktoré pracujú, keď hovoríte: „Nie!“

Hibbard inštinktívne vie, že drobná molekulárna usmievavá tvár nie je „úsmev“, pretože vie, že to nie je to, čo chce, aby jeho hypotetická UI robila. Keby niekto iný dostal nejakú inú úlohu, napríklad klasifikovať umelecké diela, mohol by mať dojem, že Mona Lisa sa samozrejme usmieva – na rozdiel od povedzme mračenia sa – hoci je to len obrázok.

A ako ilustruje prípad Terry Schiavo, technológia umožňuje nové hraničné prípady, ktorá nás vrhajú do nových, v podstate *morálnych* dilem. Ukázať UI obrázky živých a mŕtvych ľudí, ako existovali v časoch starovekého Grécka, neumožní UI urobiť *morálne* rozhodnutie, či je vypnutie podpory života Terry vražda. Táto informácia sa v databáze nenachádza ani induktívne! Terry Schiavo vytvára nové morálne otázky, odvolávajúc sa na nové morálne úvahy, na ktoré ste nemuseli myslieť, kým ste triedili fotografie živých a mŕtvych ľudí z čias starovekého Grécka. Vtedy nikto nebol na podpore života, dýchajúc, s napoly rozpusteným mozgom. Takže takéto úvahy nehrali žiadnu rolu v kauzálnom procese, ktorý ste použili na klasifikovanie tréningových údajov zo starovekého Grécka, a preto nevrhali žiaden tieň na tréningové údaje, a preto nie sú dostupné pomocou indukcie z tréningových údajov.

Čo sa týka formálneho klamu, vidím tu vystavené dve antropomorfné chyby.

Prvý klam je *podcenenie zložitosti pojmu*, ktorý vyvíjame kvôli jeho hodnote. Hranice tohto pojmu budú záležať na mnohých hodnotách a pravdepodobne dodatočnom morálnom uvažovaní, ak je hraničný prípad takého typu, aký sme predtým nevideli. Ale toto všetko sa odohráva neviditeľne v pozadí; takže sa Hibbardovi jednoducho zdá, že drobná molekulárna usmievavá tvár samozrejme nie je úsmev. A my negenerujeme *všetky* možné hraničné prípady, takže nemyslíme na všetky okolnosti, ktoré by mohli hrať rolu pri predefinovaní tohto pojmu, ale zatiaľ nehrali rolu pri jeho definovaní. Keďže ľudia podceňujú zložitosť svojich pojmov, podceňujú zložitosť indukcie pojmu z tréningových údajov. (A tiež náročnosť opísania tohto pojmu priamo – viď Skrytá zložitosť želaní.)

Druhým klamom je antropomorfný optimizmus: Keďže Bill Hibbard používa svoju inteligenciu, aby vytvoril možnosti, a plánuje dosiahnuť vysokú polohu na svojom rebríčku preferencií, je nedôverčivý voči predstave, že by superinteligencia mohla klasifikovať dovtedy nevidené drobné molekulárne usmievavé tváre ako pozitívne príklady „úsmevu“. Ako Hibbard používa pojem „úsmev“ (aby opísal požadované správanie superinteligencie), rozšírenie „úsmevu“, aby zahŕňal drobné molekulárne usmievavé tváre by bolo hodnotené príliš nízko na jeho rebríčku preferencií; bol by to *hlúpy* nápad – zo svojej podstaty hlúpy, ako vlastnosť samotného tohto pojmu – takže by to iste superinteligencia neurobila; toto je skrátka očividne *nesprávna* klasifikácia. *Superinteligencia* by určite videla, ktoré kôpky kamienkov sú správne a nesprávne.

Ale veď Priateľská UI vôbec nie je taká ťažká! Jediné, čo potrebujete, je UI, ktorá robí to, čo je *dobré*! Ach, iste, nie každá možná myseľ robí to, čo je dobré – ale v tomto prípade jednoducho *naprogramujeme* túto superinteligenciu, aby robila to, čo je *dobré*. Jediné, čo potrebujete, je neurónová sieť, ktorá uvidí pár príkladov *dobrých* vecí a nie-*dobrých* vecí, a máte klasifikáciu. Zapojte to do maximalizátora očakávaného úžitku, a hotovo!

Toto budem nazývať klamom čarovných kategórií – jednoduché malé slová, o ktorých vysvitne, že nesú všetku želanú funkcionalitu UI. Prečo nenaprogramovať šachový stroj tak, že spustíme neurónovú sieť (čiže, pohlcovač čarovných kategórií) nad množinou vyhrávajúcich a prehrávajúcich postupností šachových ťahov, aby vedela generovať „vyhrávajúce“ postupnosti? Späť v 1950-tych rokoch sa verilo, že UI bude takáto jednoduchá, ale *ukázalo sa, že to tak nie je*.

Začiatočník si myslí, že Priateľská UI je problémom *donútenia* UI, aby robila to, čo *vy* chcete, namiesto aby sa riadila svojimi vlastnými túžbami. Lenže skutočný problém Priateľskej UI je problém *komunikácie* – prenesenia hraníc kategórií, ako je „dobrý“, ktoré nemožno celkom ohraničiť pomocou žiadnych tréningových údajov, ktoré dáte UI počas jej detstva. V porovnaní s celým priestorom možností, ktoré Budúcnosť obsahuje, my *sami* sme si nepredstavili väčšinu hraničných prípadov, a museli by sme sa pustiť do plne rozvinutých morálnych debát, aby sme ich vyriešili. Aby ste vyriešili problém PUI, musíte vystúpiť von z paradigmy indukcie na ľuďmi označených trénovacích údajoch, *aj* paradigmy ľuďmi vytvorených intenzionálnych definícií.

Samozrejme, aj keby sa Hibbardovi podarilo sprostredkovať UI pojem, ktorý presne zahŕňa všetky výrazy ľudskej tváre, ktoré by Hibbard označil ako „úsmev“ a nezahŕňa žiaden výraz ľudskej tváre, ktorý by Hibbard neoznačil ako „úsmev“...

Potom by výsledná UI počas svojho detstva fungovala *napohľad* správne, dokiaľ by bola dosť slabá na to, aby dokázala generovať úsmevy iba tak, že svojim programátorom urobí radosť.

Keby UI pokročila do bodu superinteligencie a vlastnej nanotechnologickej infraštruktúry, odtrhla by vám tvár, zadrôtovala by do nej trvalý úsmev, a začala by xeroxovať.

Hlboké odpovede na takéto problémy sú mimo rozsahu tohto článku, ale je všeobecným princípom Priateľskej UI, že nemá žiadne náplaste. V roku 2004 Hibbard upravil svoj návrh tvrdením, že výrazy ľudského súhlasu by mali posilniť definíciu šťastia, a potom by šťastie malo posilňovať iné správanie. Čo, aj keby fungovalo, by akurát viedlo k tomu, že si UI naxeroxuje hordu vecí podobných-v-jej-priestore-pojmov na programátorov hovoriacich: „Áno, toto je šťastie!“ o atómoch vodíka – atómy vodíka sa robia ľahko.

Tu je odkaz na moju diskusiu s Hibbardom. To dôležité už viete.

*→

275. Skutočná väzenská dilema

Jedného dňa mi napadlo, že štandardné znázornenie Väzenskej dilemy je nesprávne.

V jadre Väzenskej dilemy je táto symetrická matica odmien:

1: S	1: P

→ http://lesswrong.com/lw/td/magical_categories/

2: S	(3, 3)	(5, 0)
2: P	(0, 5)	(2, 2)

Hráč 1 a hráč 2 si môžu každý vybrať S alebo P. Výsledný úžitok pre 1 a 2 sú prvé a druhé číslo v dvojici. Z dôvodov, ktoré budú zrejmými, „S“ znamená „spolupráca“ a „P“ znamená „podraz“.

Všimnite si, že hráč v tejto hre (berieme z pohľadu prvého hráča) má tento rebríček preferencií ohľadom výsledok: (P, S) > (S, S) > (P, P) > (S, P).

Zdalo by sa, že P dominuje S: Ak si druhý hráč vyberie S, pre vás je lepšie (P, S) než (S, S); a ak si druhý hráč vyberie P, pre vás je lepšie (P, P) než (S, P). Takže si múdro vyberiete P, a keďže matica odmien je symetrická, druhý hráč si tiež vyberie P.

Kiež by ste boli obaja menej múdri! *Obaja* dávate prednosť (S, S) pred (P, P). Čiže, obaja dávate prednosť vzájomnej spolupráci pred vzájomným podrazom.

Väzenská dilema je jedna z veľkých základných tém v teórii rozhodovania, a napísalo o nej ohromné množstvo materiálu. Čo robí drzým moje tvrdenie, že zvyčajný spôsob, ako si *predstavujeme* Väzenskú dilemu má vážnu chybu, prinajmenšom pokiaľ ste človek.

Klasické znázornenie Väzenskej dilemy je nasledujúce: ste zločinec, a vy a váš spolupáchateľ ste boli obaja dopadnutí štátnou mocou.

Nezávisle, bez možnosti spolu komunikovať, a bez možnosti neskôr zmeniť svoj názor, sa musíte rozhodnúť, či budete svedčiť proti vášmu spolupáchateľovi (P) alebo zostanete ticho (S).

Obom z vás momentálne hrozí jeden rok väzenia; svedectvo (P) odoberie jeden rok z vášho rozsudku, a pridá dva roky k rozsudku vášho spolupáchateľa.

Alebo možno vy a nejaký neznámy človek, iba raz, bez poznania histórie toho druhého a bez možnosti dodatočného zistenia, kto to bol, sa rozhodujete, či hrať S alebo P, o odmenu v dolároch zodpovedajúcu štandardnému grafu.

Ach, áno – v tej klasickej vizualizácii si máte *predstaviť, že ste dokonale sebecký*, a že vám nezáleží na vašom spolupáchateľovi, alebo na hráčovi v druhej miestnosti.

Toto posledné upresnenie robí toto klasické znázornenie, podľa môjho názoru, falošným.

Nemôžete sa vyhnúť klamu spätného pohľadu, ak dáte porote pokyny, aby sa tvárila, že nepozná skutočný výsledok množiny udalostí. A bez komplikovaného úsilia, za ktorým stojí značné poznanie, nemôže neurologicky nepoškodený človek predstierať, že je naozaj, dokonale sebecký.

Narodili sme sa so zmyslom pre spravodlivosť, česť, empatiu, súcit, a dokonca altruizmus – výsledok toho, že naši predkovia sa adaptovali na hranie *iterovanej* Väzenskej dilemy. Nedokážeme naozaj, úprimne, absolútne a naprosto dávať prednosť (P, S) pred (S, S), aj keď môžeme naprosto dávať prednosť (S, S) pred (P, P), a (P, P) pred (S, P). Predstava, že náš spolupáchateľ strávi tri roky vo väzení, nás nemôže vôbec neovplyvniť.

V tej zamknutej kobke, kde hráme jednoduchú hru pod dozorom ekonomických psychológov, nie sme naprosto a absolútne necitliví voči cudzincovi, ktorý by mohol spolupracovať. Nie sme naprosto šťastní z predstavy, že by sme mohli podraziť a cudzinec spolupracovať, čím by sme získali päť dolárov, a cudzinec nedostane nič.

Inštinktívne sa fixujeme na výsledok (S, S) a hľadáme spôsob, ako argumentovať, že by to malo byť spoločné rozhodnutie: „Ako môžeme zabezpečiť vzájomnú spoluprácu?“ je inštinktívna myšlienka. Nie: „Akým trikom môžem dosiahnuť, aby druhý hráč hral S, kým ja zahrám P pre maximálnu odmenu?“

Pre niekoho so sklonom k altruizmu alebo cti alebo spravdlivosti, nemá Väzenská dilema *naozaj* tú kritickú maticu odmien – bez ohľadu na to, aké sú *finančné* odmeny pre jednotlivcov. (S, S) > (P, S), a tá hlavná otázka je, či to ten druhý hráč vidí tiež rovnako.

A nie, nemôžete dať pokyny ľuďom, ktorých práve zasväcujete do teórie hier, že majú predstierať, že sú celkom sebeckí – rovnako ako nemôžete dávať pokyny ľuďom, ktorých práve zasväcujete do antropomofizmu, aby predstierali, že sú maximalizátori očakávaných kancelárskych spiniek.

Aby sme zostavili Skutočnú väzenskú dilemu, situácia musí vyzerať nejako takto:

Hráč 1: Ľudia, Priateľská UI, alebo iná humánna inteligencia.

Hráč 2: Nie-Priateľská UI, alebo mimozemšťan, ktorému záleží iba na triedení kamienkov.

Predstavme si, že štyri miliardy ľudí – nie celý ľudský druh, ale jeho podstatná časť – práve trpí smrteľnou chorobou, ktorú môže vyliečiť iba látka X.

Látku X však možno vyrobiť iba v spolupráci s maximalizátorom kancelárskych spiniek z inej dimenzie – látka X sa používa aj na výrobu kancelárskych spiniek. Maximalizátorovi kancelárskych spiniek záleží iba na počte kancelárskych spiniek v jeho vlastnom vesmíre, nie v našom, takže mu nemôžeme ponúknuť, že mu tam vyrobíme kancelárske spinky, ani sa vyhrážať, že ich zničíme. Nikdy predtým sme s maximalizátorom kancelárskych spiniek neinteragovali, a nikdy viac ani nebudeme.

Aj ľudstvo aj maximalizátor kancelárskych spiniek dostanú jedinú šancu zmocniť sa ďalšej časti látky X pre seba, tesne pred tým, ako spojenie medzi dimenziami skolabuje; ale proces dobývania zničí časť látky X.

Matica odmien je takáto:

	1: S	1: P
2: S	(zachránené 2 miliardy ľudských životov, získané 2 kancelárske spinky)	(+3 miliardy životov, +0 kancelárskych spiniek)
2: P	(+0 životov, +3 kancelárske spinky)	(+1 miliarda životov, +1 kancelárska spinka)

Vybral som takúto maticu odmien, aby som vytvoril pocit *rozhorčenia* pri predstave, že maximalizátor kancelárskych spiniek je ochotný vymeniť miliardy ľudských životov za pár kancelárskych spiniek. Je jasné, že maximalizátor kancelárskych spiniek by jednoducho *mal* dať všetku látku X nám; lenže maximalizátor kancelárskych spiniek nerobí to, čo by *mal*, ale jednoducho maximalizuje kancelárske spinky.

V tomto prípade *naozaj* dávame prednosť výsledku (P, S) pred výsledkom (S, S), bez ohľadu na spôsob, ako k tomu došlo. Omnoho radšej by sme žili vo vesmíre, kde boli z choroby uzdravené 3 miliardy ľudí a nevyrobili sa žiadne kancelárske spinky, než obetovať jednu miliardu ľudských životov na vytvorenie 2 kancelárskych spiniek. V prípade, ako je tento, sa nezdá *správne* spolupracovať. Nevyzerá to ani *spravodlivo* – taká veľká obeť od nás, pre taký malý zisk maximalizátora kancelárskych spiniek? A upresnime, že tento spinkový činiteľ neprežíva žiadnu bolesť ani radosť – iba dáva na výstup akcie, ktoré navigujú jeho vesmír, aby obsahoval viac kancelárskych spiniek. Spinkový činiteľ nebude prežívať žiadnu radosť zo získania kancelárskych spiniek, žiadnu bolesť zo straty kancelárskych spiniek, ani žiaden bolestivý pocit, že sme ho zradili.

Čo urobíte vtedy? Budete spolupracovať vtedy, keď naozaj, definitívne, skutočne a absolútne chcete tú najväčšiu odmenu, akú môžete získať, a ani trošičku vám v porovnaní s tým nezáleží na tom, čo sa stane tomu druhému hráčovi? Keď sa zdá *správne* podraziť ešte aj keď ten druhý hráč spolupracuje?

Takto vyzerá matica odmien *skutočnej* Väzenskej dilemy – situácia, kde (P, S) vyzerá *správnejšie* než (S, S).

Ale celý zvyšok logiky – všetko o tom, čo sa stane, ak obaja činitelia rozmýšľajú týmto spôsobom a obaja podrazia – zostáva rovnako. Pretože maximalizátorovi kancelárskych spiniek sú ľudská smrť, ľudská bolesť, a ľudský pocit zrady rovnako ľahostajné, ako sú nám kancelárske spinky. A predsa obaja dávame prednosť (S, S) pred (P, P).

Ak ste sa niekedy pýšili tým, že by ste spolupracovali vo Väzenskej dileme... alebo ste pochybovali o závere klasickej teórie hier, že „rozumná“ voľba je podraziť... čo potom poviete na horeuvedenú Skutočnú väzenskú dilemu?

PS: V skutočnosti si nemyslím, že rozumní činitelia by mali vždy podraziť v jednorazových Väzenských dilemách, kde druhý hráč bude spolupracovať vtedy, keď očakáva, že urobíte to isté. Myslím

si, že existujú situácie, kde dvaja činitelia môžu rozumne dosiahnuť (S, S) namiesto (P, P), a získať príslušné výhody.[247]

Niektoré zo svojich úvah vysvetlím, keď budeme hovoriť o Newcombovom probléme. Ale nemôžeme sa rozprávať o tom, či je možná rozumná spolupráca v tejto dileme, dokiaľ sme sa nezbavili inštinktívneho pocitu, že výsledok (S, S) je pekný alebo dobrý sám osebe. Musíme vidieť ďalej ako len po prosociálnu nálepku „vzájomná spolupráca“, ak máme pochopiť matematiku. Ak máte intuíciu, že (S, S) je z pohľadu Hráča 1 lepšie ako (P, P), ale nemáte intuíciu, že (P, S) je tiež lepšie ako (S, S), ešte si neuvedomujete, čo robí tento problém zložitým.

*→

276. Súcitné mysle

„Zrkadlové neuróny“ sú neuróny, ktoré sú aktívne aj keď nejakú činnosť robíte, aj keď tú istú činnosť sledujete – napríklad neurón, ktorý posiela signál, keď zdvihnete prst, alebo keď vidíte, ako niekto iný zdvihol prst. Takéto neuróny boli priamo namerané u primátov, a zodpovedajúce neurozobrazovacie indície sa našli u ľudí.

Možno si spomínate z môjho predchádzajúceho písania o „empatickom usudzovaní“ myšlienku, že mozgy sú také zložité, že jediný spôsob, ako ich simulovať, je donútiť podobný mozog, aby sa správal podobne. Mozog je taký zložitý, že keby sa človek pokúsil pochopiť mozog takým spôsobom, ako chápeme napríklad gravitáciu alebo autá – pozorovať celok, pozorovať časti, postaviť teóriu od základu – nedokázali by sme *vymyslieť dobré hypotézy* za celú dobu svojho života. Jediný možný spôsob, ako môžete naraziť na „Aha!“, ktoré opisuje systém taký neuveriteľne zložitý ako je Iná Myseľ, je ak natrafíte na niečo úžasne podobné tejto Inej Mysli – konkrétne svoj vlastný mozog – čo dokážete naozaj donútiť, aby sa správalo podobne, a použijete to ako hypotézu, ktorá dáva predpovede.

Toto je teda to, čo nazývam „empatia“.

A „súcit“ je potom ešte niečo navrchu tohto – usmievať sa, keď vidíte, že sa niekto iný usmieva, cítiť bolesť, keď vidíte, že niekto iný je zranený. Ide to za oblasť predpovedania, do oblasti posilňovania.

Opýtate sa: „Prečo by bezohľadný prirodzený výber robil niečo *také* pekné?“

Mohlo to začať, možno, láskou matky k svojim deťom, alebo bratskou láskou k súrodencovi. Chcete, aby žili, chcete, aby mali jedlo, iste; ale ak sa usmievate, keď sa oni usmievajú a zvíjate, keď sa oni zvíjajú, je to jednoduché nutkanie, ktoré vás vedie k pomáhaniu rozmanitými spôsobmi, v mnohých oblastiach života. Pokiaľ ste v pravekom prostredí, vaši príbuzní pravdepodobne chcú niečo, čo súvisí s úspechom v rozmnožovaní vašich príbuzných – toto je samozrejme vysvetlenie pre selekčný tlak, nie vedomý názor.

Môžete sa opýtať: „Prečo nevyvinúť abstraktnejšiu túžbu vidieť, ako niektorí ľudia, označení ako ‚príbuzní‘, dostanú čo chcú, bez toho, že by ste sami naozaj cítili to, čo cítia oni?“ Pokrčil by som plecami a povedal: „Lebo potom by ste potrebovali celú definíciu slova ‚chcieť‘ a podobne. Evolúcia si nevyberá zložitú správnu optimálnu cestu, ale stúpa nahor terénom spôsobilosti tak, ako voda tečie dole kopcom. Architektúra na zrkadlenie tam už bola, takže bol krátky krok od empatie k súcitu, a už je to hotové.“

Príbuzní – a ešte vzájomnosť; vaši spojenci v kmeni, tí, s ktorými si vymieňate láskavosti. Niečo za niečo, alebo prepracovanejšia verzia evolúcie, ktoré zohľadňuje spoločenskú povesť.

Kto je ten najimpozantnejší medzi ľuďmi? Ten najsilnejší? Ten najbystrejší? Omnoho častejšie než hociktorý z nich, myslím si, je ten, kto si dokáže zavolať najviac kamarátov.

Ako si teda urobíte veľa kamarátov?

Mohli by ste napríklad mať konkrétne nutkanie nosiť vašim spojencom jedlo, ako majú netopiere upíry – majú vo svojich kolóniách celý systém vzájomného darovania krvi. Ale je omnoho *všeobecnejšia*

247 Eliezer Yudkowsky, *Timeless Decision Theory*, Neuverejnený rukopis (Machine Intelligence Research Institute, Berkeley, CA, 2010), http://intelligence.org/files/TDT.pdf.

→ http://lesswrong.com/lw/tn/the_true_prisoners_dilemma/

motivácia, ktorá spôsobí, že si organizmus nazbiera *viac* láskavostí, ak sa usmievate vtedy, keď sa vaši vybraní priatelia usmievajú.

A aký druh organizmu sa bude vyhýbať tomu, aby naňho jeho kamaráti nahnevali, celkom všeobecne? Taký, ktorý sa zvíja, keď sa oni zvíjajú.

Samozrejme chcete byť aj schopný zabiť vybraných nepriateľov bez výčitiek svetomia – hovoríme tu predsa o ľuďoch.

Lenže... nie som si týmto istý, ale *pripadá* mi to, že súcit je medzi ľuďmi štandardne „zapnutý“. Existujú kultúry, ktoré pomáhajú cudzím... aj kultúry, ktoré jedia cudzích; otázka je, ktoré z toho si vyžaduje vyslovené prikázanie, a ktoré je štandardným ľudským správaním. Nemyslím si, že by som bol celkom šialený idealistický blázom, keď poviem, že podľa mojich znalostí antropológie, ktorých obmedzenosť priznávam, to vyzerá, že súcit je štandardne zapnutý.

Každopádne... je bolestivé, ak sa prizeráte vojne medzi dvoma stranami, a váš súcit *nebol* pre žiadnu z nich vypnutý, takže sa zvíjate, keď vidíte mŕtve dieťa bez ohľadu na to, aký je pri fotografii uvedený text; a predsa tieto dve strany voči sebe navzájom nemajú súcit, a pokračujú v zabíjaní.

Toto je teda ľudský jazyk *súcitu* – zvláštna, zložitá, hlboká implementácia vzájomnosti a pomoci. Spája mysle dokopy – nie pomocou položku vo funkcii úžitku pre „túžbu“ nejakej inej mysle, ale jednoduchšou a predsa omnoho konsekvencialistickejšou cestou zrkadlových neurónov: cítením toho, čo cíti iná myseľ, a hľadaním podobných stavov. Dokonca aj keď sa to deje pomocou pozorovania a usudzovania, a zatiaľ nie priamym prenosom neurónových informácií.

Empatia je ľudským spôsob, ako predpovedať iné mysle. Nie je to *jediný* možný spôsob.

Ľudský mozog nemožno *rýchlo* prestavať; ak sa náhle ocitnete v tmavej miestnosti, nemôžete si prerobiť zrakovú kôru na sluchovú kôru, aby ste lepšie spracovávali zvuky, dokiaľ nevyjdete, a potom by ste rýchlo prepli všetky neuróny späť, aby opäť boli zrakovou kôrou.

UI, prinajmenšom taká, ktorá funguje na niečom podobnom modernej programovacej architektúre, môže jednoducho presúvať výpočtové zdroje z jedného vlákna do druhého. Ocitneš sa v tme? Vypni zrakové vnímanie a venuj všetky operácie zvukovému; odlož starý program na disk a uvoľni pamäť, potom načítaj späť z disku, keď sa zapne svetlo.

Načo by UI potrebovala dostať *svoju* myseľ do stavu podobnému tomu, čo chce predpovedať? Stačí vytvoriť *samostatnú* inštanciu mysle – možno s odlišným algoritmom, aby lepšie simulovala tohto veľmi nepodobného človeka. Nepokúšajte sa miešať tieto údaje s vaším vlastným stavom mysle; nepoužívajte zrkadlové neuróny. Myslite na všetky riziká a zmätky, ktoré *toto* prináša!

Maximalizátor očakávanej užitočnosti – najmä taký, ktorý rozumie inteligencii na abstraktnej úrovni – má iné možnosti ako *empatiu*, keď treba pochopiť iné mysle. Tento činiteľ si nepotrebuje predstaviť, aké by to bolo byť *sám* v koži niekoho iného; môže jednoducho danú myseľ modelovať *priamo*. Hypotéza ako hocijaká iná, iba o čosi väčšia. Nemusíte sa stať svojou topánkou, aby ste pochopili topánku.

A súcit? Nuž, predstavte si, že máme do činenia s maximalizátorom očakávaných kancelárskych spiniek, ale takým, ktorý ešte nie je dosť mocný na to, aby si vo všetkom presadil svoje – musí jednať s ľuďmi, aby získal svoje kancelárske spinky. Takže tento spinkový činiteľ... modeluje týchto ľudí ako dôležité časti svojho prostredia, modeluje ich pravdepodobné reakcie na rôzne podnety, a robí veci, ktoré spôsobia, že mu ľudia budú v budúcnosti viac naklonení.

Pre maximalizátor kancelárskych spiniek sú ľudia jednoducho stroje, na ktorých možno stláčať tlačidlá. Netreba *cítiť, čo cíti ten druhý* – keby to vôbec bolo možné pri priepastnom rozdiele v internej architektúre. Ako by sa mohol maximalizátor očakávaných kancelárskych spiniek „cítiť štastne“ pri pohľade na ľudský úsmev? „Štastie“ je výraz v pravidlách učenia posilňovaním, nie v maximalizácii očakávaného úžitku. Maximalizátor kancelárskych spiniek sa necíti šťastný, keď vyrába kancelárske spinky, on si iba vyberá vždy tú činnosť, ktorá vedie k najväčšiemu počtu očakávaných kancelárskych spiniek. Maximalizátor spiniek však môže zistiť, že sa výhodné zobrazovať úsmev, keď vyrába kancelárske spinky – aby mu to pomohlo manipulovať tých ľudí, ktorí si ho označili ako priateľa.

Môže vám pripadať trochu zložité predstaviť si taký algoritmus – predstaviť si sám seba v koži niečoho, čo nefunguje tak, ako vy, a nefunguje ako žiaden režim, do ktorého sa váš mozog dokáže prepnúť.

Môžete svoj mozog prepnúť do režimu nenávidenia nepriateľa, ale ani to nie je správne. Aby ste si predstavili, ako naozaj *nesúcitná* myseľ vidí človeka, musíte si sami seba predstaviť ako užitočný stroj, na ktorom sú rôzne páky. Nie ako stroj v tvare človeka, lebo voči tomu máme inštinkty. Jednoducho ako pílu alebo niečo také. Niektoré páky spôsobia, že zo stroja vypadnú peniaze, iné páky môžu spôsobiť, že vystrelí guľku. Tento stroj má trvalý vnútorný stav a vy musíte páky potiahnuť v správnom poradí. Ale stále je to iba zložitý kauzálny systém – nie je v ňom nič vnútorne myšlienkové.

(Aby ste pochopili *nesúcitiaci* optimalizačný proces, odporučil by som študovať prirodzený výber, ktorý sa neunúva tlmiť bolesť smrteľne zraneným a zomierajúcim tvorom, hoci ich bolesť už neslúži žiadnemu rozmnožovaciemu účelu, pretože ani tlmenie bolesti by neslúžilo žiadnemu rozmnožovacieho účelu.)

Preto som uviedol „súcit“ ešte pred „nudou“ vo svojom zozname vecí, ktoré by museli mať mimozemšťania, aby boli aspoň trochu, pardon za výraz, sympatickí. Nie je nemožné, že by súcit mohol existovať vo významnom zlomku všetkých vyvinutých mimozemských inteligentných druhov; zrkadlové neuróny vyzerajú ako ten druh veci, ktorý ak sa stal raz, *môže* sa stať znovu.

Nesúcitní mimozemšťania môžu byť obchodnými partnermi – alebo nie, veď hviezdy a podobné suroviny sú prakticky rovnaké v celom vesmíre. Mohli by sme s nimi vyjednať mierové dohody, a oni by ich mohli dodržiavať z vypočítavého strachu z odvety. Mohli by sme s nimi dokonca spolupracovať vo Väzenskej dileme. Ale nikdy by sme sa nimi nemohli kamarátiť. Nikdy by nás nevideli ako nič iné než prostriedok na dosiahnutie nejakého cieľa. Nepreliali by kvôli nám jedinú slzu, ani by sa neusmiali nad našou radosťou. A rovnako by zaobchádzali aj s ostatnými zo svojho vlastného druhu, a nemali by žiaden pocit, že im vďaka tomu niečo dôležité chýba.

Takíto mimozemšťania by boli *varelse*, nie *ramen* – ten druh mimozemšťanov, s ktorým nemôžete mať žiaden osobný vzťah a ani nemá zmysel sa oň pokúšať.

*→

277. Veľká výzva

Existuje množina proroctiev, ktoré znejú: „V Budúcnosti budú všetku prácu robiť stroje. Všetko bude automatizované. Dokonca aj takú práca, ktorú dnes považujeme za ‚intelektuálnu‘, ako inžinierstvo, budú robiť stroje. My budeme sedieť a vlastniť kapitál. Nikdy viac už nebudeme musieť pohnúť prstom.“

Ale nebudú sa potom ľudia nudiť?

Nie; budú sa hrať počítačové hry – nie ako tie *naše*, samozrejme, ale omnoho vyspelejšie a zaujímavejšie.

Ale počkajte! Ak si kúpite modernú počítačovú hru, zistíte, že obsahuje nejaké úlohy, ktoré sú – na toto nie je žiadne pekné slovo – *namáhavé*. (Dokonca by som povedal „ťažké“, ak si rozumieme, že hovoríme o niečom, čo zaberie 10 minút, nie 10 rokov.)

Takže v budúcnosti budeme mať programy, ktoré nám *pomôžu* hrať túto hru – prevezmú to za nás, ak sa v hre zasekneme, alebo iba začneme nudiť; alebo aby sme mohli hrať hry, ktoré by pre nás inak boli príliš náročné.

Ale nie je toto trochu plytvanie námahou? Prečo by jeden programátor pracoval na tom, aby bola hra ťažšia, a druhý programátor na tom, aby bola hra ľahšia? Prečo tú hru *rovno neurobiť* ľahkou? Keďže tú hru hráte, aby ste získali zlato a body skúsenosti, ľahšia hra vám umožní získať viac zlata za jednotku času: hra bude zábavnejšia.

Takže toto je úplný cieľ proroctva o technologickom pokroku – iba pozeranie na obrazovku, ktorá hovorí „VYHRAL SI“, naveky.

→ http://lesswrong.com/lw/xs/sympathetic_minds/

A možno aj na *toto* si postavíme robota.

Potom čo?

Svet strojov, ktoré robia *všetku* prácu – no, nechcem povedať, že je „analogický kresťanskému nebu“, pretože nie je nadprirodzený; je to niečo, čo by sa v princípe dalo uskutočniť. Náboženskými analógiami sa pri obviňovaní príliš ľahko ohadzuje... Ale, bez naznačovania ľubovoľných ďalších podobností, by som povedal, že to znie analogicky v tom zmysle, že večná lenivosť „znie ako dobrá správa“ pre vaše terajšie ja, ktoré stále musí pracovať.

A čo sa týka hrania hier ako náhrady – čo *je* to počítačová hra, ak nie umelá práca? Nie je to zbytočný krok? (Ak sa zamyslíme nad počítačovými hrami v ich dnešnej forme ako nad prácou, majú rôzne stránky, ktoré zmenšujú stres a zvyšujú zapojenie; ale majú aj svoju cenu v podobe umelosti a osamelosti.)

Niekedy si myslím, že futuristické ideály vyjadrené slovami „zbaviť sa práce“ by bolo lepšie preformulovať ako „odstrániť málo kvalitnú prácu, aby bolo miesto na vysoko kvalitnú prácu“.

Existuje veľká skupina cieľov, ktoré sa nehodia ako dlhodobý zmysel života, pretože ich môžete naozaj dosiahnuť, a potom ste hotoví.

Ak sa na to pozrieme z iného uhla, ak hľadáme primeraný dlhodobý zmysel života, mali by sme hľadať ciele, o ktoré je dobré sa *usilovať*, nie iba dobré ich *dosiahnuť*.

Alebo aby sme to povedali trochu menej paradoxne: Mali by sme sa pozerať na hodnotenia 4D stavov, namiesto 3D stavov. Hodnotiť prebiehajúce proces, namiesto „nech má vesmír vlastnosť P a potom hotovo“.

Tu sa opäť oplatí citovať Timothyho Ferrisa: Aby sme našli šťastie, „otázka, ktorú by ste si mali položiť nie je: ‚Čo chcem?‘ ani ‚Čo sú moje ciele?‘ ale ‚Čo by ma bavilo?‘“

Mohli by ste povedať, že pre dlhodobý zmysel života potrebujeme hry, ktoré je zábavné *hrať*, a nie iba *vyhrať*.

Pripomínam – niekedy *chcete* vyhrať. Existujú legitímne ciele, kde vyhrať znamená všetko. Ak hovoríme napríklad o liečbe rakoviny, potom utrpenie, ktoré zažíva čo len jediný pacient s rakovinou prevyšuje všetku zábavu, ktorú by ste mohli mať pri riešení jeho problému. Ak dvadsať rokov pracujete na vytvorení lieku na rakovinu vlastným úsilím, učíte sa nové poznatky a nové zručnosti, robíte si priateľov a spojencov – a potom vám nejaká mimozemská inteligencia ponúkne liek na rakovinu na striebornom podnose za tridsať dolárov – potom držte hubu a vezmite to.

Ale „vyliečiť rakovinu“ je problém typu 3D výroku: chcete, aby sa výrok „rakovina neexistuje“ prepol z hodnoty Nepravda v súčasnosti na hodnotu Pravda v budúcnosti. Dôležitosť tohto cieľa výrazne prevyšuje samotnú cestu; vy tam nechcete *putovať*, vy tam skrátka chcete *byť*. Existuje mnoho *legitímnych* cieľov tohto druhu, ale nie sú vhodné na dlhodobú zábavu. „Vyliečiť rakovinu!“ je aktivita hodnotná, aby sme ju nasledovali tu a teraz, ale nie uveriteľný budúci cieľ galaktických civilizácií.

Prečo by tento „hodnotný prebiehajúci proces“ mal byť procesom *snahy robiť veci* – prečo nie procesom pasívneho vnímania, ako buddhistické nebo?

Priznám sa, že si nie som celkom istý, ako nastaviť „pasívne vnímajúcu“ myseľ. Ľudský mozog je *zostavený*, aby robil rôzne druhy vnútornej práce, z ktorých sa skladá aktívna inteligencia; aj keby ste ležali na posteli a nevynakladali nijakú konkrétnu myšlienkovú námahu, myšlienky, ktoré idú vaším mozgom sú aktivitami oblastí mozgu, ktoré sú zostavené, aby, veď viete, *riešili problémy*.

Koľko veľa z ľudského mozgu by ste mohli odstrániť, *okrem* centier radosti, a stále si zachovať subjektívny zážitok radosti?

Toto nejdem rozoberať. Budem sa držať omnoho jednoduchšej odpovede: „V skutočnosti by som *nechcel* byť pasívnym pozorovateľom.“ Keby som *chcel* nirvánu, mohol by som skúsiť zistiť, ako dosiahnuť túto nemožnú vec. Ale akonáhle si odmyslím, že mi Buddha hovorí, že nirvána je konečný cieľ existencie, vyzerá nirvána skôr ako „niečo, čo znie ako dobrá správa, keď to počujete po prvýkrát“ alebo „ideologická viera v túžbu“ než, viete, niečo, čo by som naozaj *chcel*.

Dôvod, prečo mám vôbec nejakú myseľ, je že ma prirodzený výber zostavil, aby som *robil* veci – riešil rôzne druhy problémov.

„Pretože je to v ľudskej povahe“ nie je explicitné ospravedlnenie hocičoho. Existuje ľudská povaha v zmysle: to, čo sme; a existuje ľudská povaha v zmysle: to, čo si ako ľudia želáme, aby sme boli.

Ale ja *nechcem* zmeniť svoju povahu smerom k pasívnejšiemu objektu – čo *je* zdôvodnenie. Šťastná machuľa *nie je* to, čo si ako človek želám, aby som sa stal.

Už som predtým tvrdil, že mnohé hodnoty si vyžadujú aj subjektívne šťastie, aj vonkajšie objekty tohto šťastia. Že môžete legitímne mať funkciu úžitku, ktorá hovorí: „Záleží mi na tom, či osoba, ktorú milujem, je skutočný človek alebo iba vysoko realistický nevedomý diskusný robot, *dokonca aj keď to neviem*, pretože to čo si cením nie je môj vlastný stav mysle, ale vonkajšia skutočnosť.“ Potrebujete teda aj zážitok lásky, aj skutočnú milujúcu osobu.

Podobne môžete mať hodnotné aktivity, ktoré vyžadujú aj skutočnú výzvu a skutočnú námahu.

Keď pretekáte na trati, záleží na tom, že sú ostatní pretekári skutoční, a že máte skutočnú šancu vyhrať alebo prehrať. (Nehovoríme tu o fyzikálnom determinizme, ale či nejaký vonkajší optimalizačný proces explicitne vybral, že vy vyhráte tento pretek.)

A záleží na tom, že pretekáte pomocou svojej vlastnej zručnosti behať a svojej vlastnej sily vôle, a nie iba stlačíte tlačidlo s nápisom „Vyhraj“. (Hoci, keďže ste si nikdy nenavrhovali vlastné svaly na nohách, pretekáte sa s použitím sily, ktorá nie je vaša. Preteky medzi robotickými autami sú čírejšou súťažou medzi ich návrhármi. Je veľa miesta na zlepšenie ľudskej situácie.)

A záleží na tom, že to zažívate vy, vedomá bytosť. (Namiesto nejakého nevedomého procesu, ktorý by vykonával schematickú napodobeninu preteku, biliónkrát za sekundu.)

Musí tam byť skutočné úsilie, skutočné víťazstvo, a skutočný zážitok – cesta, cieľ, a putujúci.

*→

278. Vážne príbehy

Každá Utópia, ktorá bola kedy vytvorená – vo filozofii, fikcii alebo náboženstve – bola do nejakej miery miestom, kde by ste *nechceli naozaj* žiť. Nie som jediný, kto si túto dôležitú vec všimol: George Orwell povedal zhruba to isté v článku „Prečo socialisti neveria v zábavu“ a predpokladám, že mnohí iní to povedali pred ním.

Ak čítate knihy o tom, Ako Písať – a existuje *veľa* kníh o tom, Ako Písať, pretože úžasne veľa autorov kníh si myslí, že vie niečo o písaní – tieto knihy vám povedia, že príbeh musí obsahovať „konflikt“.

Presnejšie, tie *miernejšie* poučné knihy vám povedia, že príbehy obsahujú „konflikt“. Niektorí autori však hovoria otvorenejšie.

„Príbehy sú o bolesti druhých ľudí.“ Orson Scott Card.

„Každá scéna musí končiť katastrofou.“ Jack Bickham.

V dobe mojej mladistvej hlúposti som považoval za samozrejmosť, že *spisovatelia* majú výnimku z hľadania skutočnej Eutópie, pretože keby ste postavili *bezchybnú* Utópiu... aké príbehy by ste mohli písať, ktoré by sa tam odohrávali? „Kde bolo, tam bolo, žili šťastne až naveky.“ Načo by to bolo dobré, keby sa spisovateľ vedeckej fantastiky pokúsil vykresliť pozitívnu Singularitu, keď pozitívna Singularita by bola...

...koniec všetkých príbehov?

Vyzeralo to ako rozumný rámec, pomocou ktorého možno skúmať literárny problém Utópie, ale niečo na tomto konečnom závere vyvolávalo tichú, nenechavú pochybnosť.

V tej dobe som si predstavoval UI ako niečo podobné bezpečnému džinovi, ktorý spĺňa želania jednotlivcov. Takže tento záver dával istý zmysel. Keby existoval problém, skrátka by ste si želali, aby zmizol, nie? Čiže – žiaden príbeh. Ignoroval som teda túto tichú, nenechavú pochybnosť.

Omnoho neskôr, keď som došiel k záveru, že bezpečný džin nie je taký dobrý nápad, sa v spätnom pohľade zase zdalo, že „žiadne príbehy“ by mohol byť užitočný indikátor. V tomto konkrétnom prípade

→ http://lesswrong.com/lw/ww/high_challenge/

„nedokážem si predstaviť jediný príbeh, ktorý by som o takomto scenári *chcel čítať*“ ma veru mohlo nasmerovať k dôvodu „nechcel by som *naozaj žiť* v takomto scenári“.

Prehltol som teda svoj vycvičený odpor proti luddizmu a teodícei, a *skúsil* som sa nad týmto argumentom aspoň zamyslieť:

- Svet, v ktorom sa nikdy nič nepokazí, kde nikto nikdy nezažije žiadnu bolesť ani smútok, je svet, ktorý neobsahuje žiadne príbehy, ktoré by sa oplatilo čítať.
- Svet, o ktorom by ste nechceli čítať, je svet, v ktorom by ste nechceli žiť.
- Do každého eudaimonického života musí patriť trochu bolesti. QED.

V istom zmysle je jasné, že *nechceme* žiť také životy, aké sú zobrazené vo väčšine príbehov, ktoré ľudia zatiaľ napísali. Predstavte si tie naozaj veľké príbehy, tie, ktoré sa stali legendami, lebo boli tým najlepším z najlepšieho vo svojom žánri: *Ilias*, *Romeo a Júlia*, *Krstný otec*, *Watchmen*, *Planescape: Torment*, druhá séria *Buffy zabíjačky upírov*, alebo *ten záver v Tsukihime*. Existuje jediný príbeh v tomto zozname, ktorý *nie je* tragický?

Zvyčajne dávame prednosť radosti pred bolesťou, veselosti pred smútkom, a životu pred smrťou. Zdá sa však, že sa radšej vžívame do trpiacich, smutných, mŕtvych postáv. Naše príbehy o šťastnejších ľuďoch *nie sú vážne*, nie sú dosť umelecky dobré, aby si zaslúžili chválu – ale prečo teda selektívne chválime príbehy, ktoré obsahujú nešťastných ľudí? Máme z toho nejaký skrytý úžitok? Je to hlavolam, nech sa na to dívate z hociktorej strany.

Keď som bol dieťa, nedokázal som písať vedeckú fantastiku, pretože som písal tak, aby sa mojim postavám darilo *dobre* – tak ako som chcel, aby sa v skutočnom živote darilo mne. Z čoho ma vyliečil Orson Scott Card: *Ach*, povedal som si, *tak toto som robil zle, moje postavy netrpia.* Ale ani potom som si neuvedomil, že mikroštruktúra zápletky funguje rovnako – dokiaľ mi Jack Bickham nepovedal, že každá scéna musí končiť katastrofou. Ja som sa snažil vytvárať problémy a *riešiť* ich, namiesto aby som ich *zhoršoval*...

Príbeh jednoducho *neoptimalizujete* tak, ako optimalizujete skutočný život. *Najlepší* príbeh a *najlepší* život budú vytvorené podľa odlišných kritérií.

V skutočnom svete ľudia dokážu žiť značnú dobu aj bez väčšej katastrofy a stále sa zdá, že sa im darí celkom dobre. Kedy vás naposledy zastrelil úkladný vrah? Dosť dávno, však? Pripadá vám váš život vďaka tomu prázdnejší?

Ale na druhej strane...

Z nejakého zvláštneho dôvodu, keď sa autori stanú príliš starými alebo príliš úspešnými, vrátia sa do stavu môjho detstva. Ich príbehy začnú ísť *správne*. Predstavú svojim postavám robiť hrozné veci, čoho dôsledkom je, že začnú robiť hrozné veci svojim čitateľom. Zdá sa, že je to bežná súčasť Syndrómu Starého Spisovateľa. Mercedes Lackey, Laurell K. Hamilton, Robert Heinlein, dokonca aj Orson Scott Card – všetci tak skončili. Zabudli, ako ubližovať svojim postavám. Neviem, prečo.

A keď si prečítate príbeh od Starého Spisovateľa alebo od čistého nováčika – príbeh, kde veci skrátka idú *neúnavne správne* jedna za druhou – kde hlavný hrdina porazí superzloducha lusknutím prstov, alebo ešte horšie, pred záverečnou bitkou sa superzloduch *vzdá a ospravedlní sa a opäť sú z nich priatelia*...

Je to ako škrípanie nechtov po tabuli pri začiatku vašej chrbtice. Ak ste v skutočnosti takýto príbeh nikdy nečítali (alebo horšie, nenapísali), potom sa považujte za šťastných.

Táto kvalita škrípania nechtov – premiestnila by sa z príbehu do skutočného života, keby ste skúšali žiť skutočný život bez jedinej kvapky dažďa?

Jedna odpoveď by mohla byť, že to, čo príbeh naozaj potrebuje, nie je „katastrofa“ ani „bolesť“ a dokonca ani „konflikt“, ale jednoducho *úsilie*. To je problém s príbehmi typu Mary Sue, že v nich nie je dosť úsilia, ale *bolesť* v skutočnosti nepotrebujú. Toto by sa možno dalo otestovať.

Alternatívna odpoveď by mohla byť, že *teraz* hovoríme o transhumanistickej verzii Teórie Zábavy. Môžeme teda odpovedať: „Upravte mozgy tak, aby sme odstránili tento pocit škrípajúcich nechtov“, pokiaľ neexistuje nejaký dôvod, prečo si ho nechať. Ak je pocit škrípajúcich nechtov nejaká zbytočná náhodná porucha, ktorá prekáža Utópii, odstráňte ju.

Možno by sme *mali*. Možno všetky Veľké Príbehy sú tragédie, pretože... no...

Raz som čítal, že v komunite BDSM je „intenzívny pocit“ eufemizmom pre bolesť. Keď som si to prečítal, napadlo mi, že tak, ako sú teraz ľudia zostavení, je skrátka *jednoduchšie* vytvárať bolesť než radosť. Hovorím tu síce trochu mimo svojej odbornosti, ale predpokladám, že vysoko talentovaný a skúsený sexuálny umelec by musel pracovať celé hodiny, aby vytvoril *dobrý* pocit rovnako intenzívny ako je bolesť jedného silného kopanca do rozkroku – ktorý zvládne sa sekundu aj nováčik.

Keď som skúmal život kňaza a proto-racionalistu Friedricha Spee von Langenfelda, ktorý počúval priznania obvinených bosoriek, vyhľadal som si niektoré z nástrojov, ktoré sa používali na vytvorenie priznania. Neexistuje žiaden bežný spôsob, ako by ste mohli dosiahnuť, aby sa človek cítil tak *dobre*, ako tieto nástroje spôsobia, aby vás bolelo. Nie som si istý, či by to dokázali dokonca aj drogy, hoci moja skúsenosť s drogami je rovnako nulová ako s mučením.

Na tomto je niečo nevyvážené.

Áno, ľudia sú vo svojom plánovaní príliš optimistickí. Keby nás straty netrápili viac než zisky, boli by sme dávno na mizine, ako sme dnes zostavení. Experimentálne pravidlo hovorí, že stratiť niečo želané – 50 dolárov, kávovú šálku, hocičo – bolí asi 2 až 2,5-krát toľko ako ekvivalentný zisk.

Ale toto je ešte hlbšia nerovnováha. Rozdiel medzi úsilím na vstupe a intenzitou na výstupe medzi sexom a mučením nie je púhy činiteľ 2.

Ak niekto chce hľadať vnemy – v tomto svete, ako sú ľudia teraz zostrojení – nie je prekvapujúce, že dôjde k bolesti zmiešanej s potešením ako k zdroju *intenzity* tohto kombinovaného zážitku.

Kiež by ľudia boli zostavení inak, takže by ste dokázali vytvárať radosť rovnako intenzívnu a v rovnako mnohých odlišných odtieňoch ako bolesť! Kiež by ste mohli, s rovnakou tvorivosťou a úsilím ako mučiteľ Inkvizície dosiahnuť, aby sa niekto cítil tak *dobre*, ako sa obete Inkvizície cítili *zle*...

Ale, čo je potom analogické potešenie, ktoré cítime ako také dobré? Obeť šikovného mučenia urobí čokoľvek, aby zastavila bolesť a čokoľvek, aby sa neopakovala. Existuje ekvivalentná radosť, ktorá prehluší všetko požiadavkou, aby pokračovala a opakovala sa? Ak majú ľudia silnejšiu vôľu znášať radosť, je to naozaj rovnaká radosť?

Existuje ďalšie pravidlo písania, ktoré hovorí, že príbehy musia *kričať*. Ľudský mozog je veľmi vzdialený od tých vytlačených písmen. Každá udalosť a pocit sa musia odohrávať pri desaťnásobku prirodzenej hlasitosti, aby mali vôbec nejaký dopad. Nesmiete sa snažiť, aby sa vaše postavy správali alebo cítili *realisticky* – najmä nesmiete verne reprodukovať svoje vlastné minulé zážitky – pretože *bez preháňania* budú príliš tiché na to, aby vystúpili zo stránok.

Možno sú všetky Veľké Príbehy tragédie, pretože šťastie nedokáže dosť nahlas kričať – na ľudského čitateľa.

Možno toto treba napraviť.

A keby sa to napravilo... zostalo by stále nejaké využitie pre bolesť alebo smútok? Aspoň na *spomienku* na smútok, keby už všetky veci boli také dobré ako len môžu byť, a každá liečiteľná choroba už vyliečená?

Môžete jednoducho rovno vymazať bolesť? Alebo odstránenie starej dolnej hranice funkcie úžitku akurát vytvorí novú dolnú hranicu? Bude každá radosť menšia než 10 000 000 hedónov novou neznesiteľnou bolesťou?

Ľudia, ako sú zostavení teraz, asi majú sklon k úprave hedonickej mierky. Niekto, kto si pamätá ako hladoval, si bude vážiť krajec chleba viac než niekto, kto nikdy nepoznal nič iné ako koláče. Toto bola hypotéza Georgea Orwella, prečo je Utópia nemožná v literatúre aj v skutočnosti:[248]

> Zdalo by sa, že ľudia nedokážu opísať, a možno si ani predstaviť, šťastie, okrem kontrastu... Neschopnosť ľudstva predstaviť si šťastie okrem úľavy, či už od námahy alebo bolesti, predstavuje pre socialistu vážny problém. Dickens dokáže opísať, ako sa chudobou sužovaná rodina vrhne na pečenú hus, a môže ich ukázať ako šťastných; na druhej strane sa

248 George Orwell, „Why Socialists Don’t Believe in Fun,“ Tribune (December 1943).

zdá, že obyvatelia dokonalých vesmírov nemajú žiadnu spontánnu veselosť a zvyčajne je trochu odporné mať s nimi do činenia.

Pre maximalizátora očakávaného úžitku nemá zmysel zmeniť škálu funkcie úžitku tak, že pripočíta bilión ku každému výsledku – je to doslova tá istá funkci úžitku, ako matematický objekt. Funkcia úžitku opisuje *relatívne* intervaly medzi výsledkami; to je to, čo je, matematicky povedané.

Lenže ľudský mozog má rôzne neurónové obvody pre pozitívnu spätnú väzbu a negatívnu spätnú väzbu, a rôzne varianty pozitívnej a negatívnej spätnej väzby. Dnes existujú ľudia, ktorí „trpia“ vrodenou analgéziou – absolútnou neprítomnosťou bolesti. Nikdy som nepočul o tom, že by im *nedostatočná radosť* pripadala neznesiteľná.

Ľudia s vrodenou analgéziou sa musia často a starostlivo kontrolovať, aby videli, či sa neporezali alebo si nepopáli prst. Bolesť má v dizajne ľudskej mysle svoj účel...

Ale to neznamená, že neexistuje žiadna alternatíva, ktorá by mohla poslúžiť rovnakému účelu. Mohli by ste vymazať bolesť a nahradiť ju *nutkaním nerobiť isté veci*, ktorému by chýbala tá neznesiteľne subjektívna kvalita bolesti? Nepoznám celý Zákon, ktorý tu vládne, ale tipol by som si, že áno, mohli; mohli by ste nahradiť túto stránku seba samého niečím podobnejším maximalizátoru očakávaného úžitku.

Mohli by ste vymazať ľudský sklon meniť škálu potešenia – vymazať zvykanie si, takže by každá nová pečená hus chutila rovnako lahodne ako predchádzajúca? Tipol by som, že mohli. Toto vedie nebezpečne blízko vymazania Nudy, ktorá je spolu so Súcitom absolútne neodmysliteľná... ale povedať, že staré riešenie zostáva rovnako príjemné, ešte neznamená povedať, že stratíte chuť hľadať nové a lepšie riešenia.

Môžete urobiť, aby každá pečená hus chutila rovnako lahodne, ako by chutila v kontraste k hladu, aj keby ste nikdy nehladovali?

Mohli by ste zabrániť tomu, aby sa bolesť zo zrnka prachu v oku stala novým mučením, ak ste doslova *nikdy nezažili* nič *horšie* než zrnko prachu dráždiace vaše oko?

Takéto otázky začínajú presahovať moje chápanie Zákona, ale tipol by som, že odpoveď je: áno, dá sa to. Mám skúsenosť s podobnými vecami, že akonáhle sa naučíte Zákon, zvyčajne dokážete vidieť, ako urobiť čudne vyzerajúce veci.

Pokiaľ viem alebo si dokážem tipnúť, David Pearce (*Hedonistický imperatív*) má pravdepodobne pravdu ohľadom *realizovateľnosti*, keď hovorí:[249]

> Nanotechnológia a genetické inžinierstvo zrušia utrpenie všetkého vedomého života. Tento projekt zrušenia je veľmi ambiciózny, ale technicky možný. Je to zároveň inštrumentálne rozumné a morálne súrne. Metabolické dráhy bolesti a nespokojnosti sa vyvinuli preto, lebo slúžili spôsobilosti našich génov v pravekom prostredí. Budú nahradené iným druhom neurónovej architektúry – motivačným systémom založeným na dedičných stupňoch blaženosti. Stavom úžasného blahobytu je súdené stať sa geneticky naprogramovanou normou duševného zdravia. Predpovedáme, že posledný nepríjemný zážitok na svete bude presne datovaná udalosť.

Je toto... to, čo *chceme*?

Jednoducho zotrieť poslednú slzu a hotovo?

Existuje nejaký dobrý dôvod *nerobiť* to, okrem skreslenia status quo a hŕstky ošúchaných racionalizácií?

Aká by bola alternatíva? Alebo alternatívy?

Nechať veci tak, ako sú? Samozrejme nie. Tento svet nenavrhol žiaden Boh; nemáme dôvod myslieť si, že je presne optimálny v ľubovoľnom smere. Ak tento svet neobsahuje priveľa bolesti, potom by jej musel obsahovať primálo, a to vyzerá nepravdepodobne.

Ale možno...

Mohli by ste odstrihnúť iba tie *neznesiteľné* časti bolesti?

---

249 David Pearce, *The Hedonistic Imperative*, http://www.hedweb.com/, 1995.

Zbavte sa Inkvizície. Ponechajte ten druh bolesti, ktorá vám hovorí, aby ste nestrkali prst do ohňa, alebo bolesť, ktorá vám hovorí, že ste nemali strčiť kamarátovi prst do ohňa, alebo dokonca aj bolesť z rozchodu s milencom.

Pokúste sa zbaviť toho druhu bolesti, ktorý *dusí a ničí* myseľ. Alebo prerobte mysle tak, aby sa ťažšie ničili.

Mohli by ste mať svet, kde by existovali zlomené nohy, dokonca aj zlomené srdca, ale nie zlomení *ľudia*. Žiadne detské sexuálne zneužívanie, ktoré vytvára ďalších násilníkov. Žiadni ľudia ubití únavou a vyčerpávajúcimi drobnými nepríjemnosťami až do bodu, keď uvažujú o samovražde. Žiadne náhodné nezmyselné nekonečné žiale ako hladovanie alebo AIDS.

A ak aj zlomená noha stále vyzerá príliš strašne...

Báli by sme sa bolesti menej, keby sme boli silnejší, keby naše každodenné životy už nevyčerpávali tak veľa z našich rezerv?

To by teda bola jedna alternatíva k Pearceovmu svetu – ak existujú ešte ďalšie alternatívy, nerozmýšľal som o nich podrobne.

Mohli by ste to nazvať cestou odvahy – myšlienka je, že keď odstránite ten ničivý druh bolesti a posilníte ľudí, to, čo zostane, by nemalo byť *také* strašné.

Svet, kde existuje smútok, ale nie masívny systematický *nezmyselný* smútok, aký vidíme vo večerných správach. Svet, kde bolesť, ak nie je odstránená, prinajmenšom *neprevažuje radosť*. O takom svete by ste mohli písať príbehy, a oni by dokázali čítať naše príbehy.

Mám sklon byť pomerne konzervatívny ohľadom predstavy vymazania veľkých častí ľudskej povahy. Nie som si istý, koľko veľa väčších častí môžete vymazať, kým sa tá vyvážená, konfliktná, dynamická štruktúra nezosype na niečo jednoduchšie, ako je maximalizátor očakávanej radosti.

A tak sa priznávam, že sa mi páči táto cesta odvahy.

Ale zase, nemám skúsenosť ani s jedným z toho.

Možno sa iba *bojím* sveta takéto odlišného ako je analgézia – nebol by to ironický dôvod, prečo kráčať „cestou odvahy“?

Možno mi cesta odvahy iba pripadá ako *menšia zmena* – možno mám jednoducho problém vžiť sa cez väčšiu priepasť.

Lenže „zmena“ je pohyblivý cieľ.

Keby ľudské dieťa vyrástlo v *menej* bolestivom svete – keby nikdy nežilo vo svete, kde je AIDS alebo rakovina alebo otroctvo, a tak by nepoznalo tieto veci ako zlo, ktoré *bolo triumfálne odstránené* – a tak by nemalo pocit, že už je „hotovo“, alebo že sa svet „už dosť zmenil“...

Čo by mu zabránilo urobiť ďalší krok, a skúsiť odstrániť neznesiteľnú bolesť zlomeného srdca, keď niekoho prestane milovať jeho milenec?

A potom čo? Existuje bod, kde *Romeo a Júlia* skrátka vyzerajú stále menej a menej relevantne, viac a viac ako relikvia nejakého vzdialeného zabudnutého sveta? Príde nejaký bod v transhumánnom putovaní, kde celá táto vec s obvodmi negatívneho posilňovania nebude vyzerať ako nič iné než zbytočná opica, z ktorej sa treba prebrať?

A ak áno, má nejaký zmysel *odkladať* tento posledný krok? Alebo by sme mali skrátka odhodiť svoj strach a... odhodiť svoj strach?

Neviem.

* →

279. Hodnota je krehká

Keby som si mal vybrať jednu vetu, ktorá sa *opiera* o obsah *Overcoming Bias* viac než hocijaká iná, akú som napísal, tá veta by bola:

→ http://lesswrong.com/lw/xi/serious_stories/

Ľubovoľná Budúcnosť, ktorú ***nevytvaroval*** *systém cieľov s podrobnou spoľahlivou dedičnosťou od ľudskej morálky a metamorálky, nebude obsahovať takmer nič hodnotné.*

„Nuž," povie niekto, „možno podľa tvojich provinčných *ľudských* hodnôt by sa *tebe* nepáčila. Ale ja si viem ľahko predstaviť galaktickú civilizáciu plnú činiteľov, ktorí vôbec nie sú takí ako *ty*, a predsa nachádzajú veľkú hodnotu a záujem vo svojich *vlastných* cieľoch. A to je podľa mňa v poriadku. Nie som taký bigotný ako ty. Nechajme Budúcnosť ísť svojou vlastnou cestou, a nepokúšajme sa ju naveky pripútať k smiešne primitívnym predsudkom tlupy štvornohých Mäkkých Vecí..."

Milý priateľ, ja *nemám problém* s predstavou budúcich civilizácii, ktoré budú značne iné ako tá naša... plných zvláštnych bytostí, ktoré zďaleka nevyzerajú ako ja, a ani si ma nevedia predstaviť... hľadajúcich radosti a zážitky, do ktorých sa nedokážem vžiť... obchodujúce na trhu s nepredstaviteľnými statkami... spájajúcimi sa za nepochopiteľnými zámermi... ľuďmi, ktorých životné príbehy by som nikdy nepochopil.

Takto bude Budúcnosť vyzerať, ak veci dopadnú *dobre*.

Ak sa reťaz dedičnosti od ľudskej (meta)morálky pretrhne, Budúcnosť *nebude* vyzerať takto. *Neskončí* ako čarovne, úžasne nezrozumiteľná.

S veľmi vysokou pravdepodobnosťou skončí ako napohľad *nudná*. Nezmyselná. Niečo, za čoho stratou by ste nesmútili.

Vidieť toto ako samozrejmosť, to si vyžaduje ohromné množstvo vysvetlenia základov.

Nebudem tu teraz prechádzať *všetkými* pointami a spletitými cestami argumentov, lebo to by sme sa vrátili späť k 75 % mojich článkov na *Overcoming Bias*. To len aby som poznamenal, koľko *veľa* rôznych vecí musíme poznať, aby sme obmedzili výslednú odpoveď.

Vezmite si tú neuveriteľne dôležitú ľudskú hodnotu „nudy" - našej túžby, nerobiť „to isté" znovu a znovu a znovu. Viete si predstaviť myseľ, ktorá by obsahovala *takmer* celú špecifikáciu ľudskej hodnoty, takmer všetku morálku a metamorálku, ale vynechala by *iba túto jednu vec*...

...a tak by strávila zvyšok času, až do najvzdialenejších koncov svojho svetelného kužeľa, prehrávaním si jediného vysoko optimalizovaného zážitku, znovu a znovu a znovu.

Alebo si predstavte myseľ, ktoré by obsahovala takmer celú špecifikáciu toho, aké druhy pocitov majú ľudia najradšej – okrem myšlienky, že tieto pocity majú dôležité *vonkajšie referenty*. Táto myseľ by sa teda skrátka *cítila*, akoby urobila dôležitý objav, *cítila*, akoby našla dokonalého milenca, *cítila*, akoby pomohla priateľovi, ale v skutočnosti by nič z toho *nerobila* – stala by sa sama sebe strojom na zážitky. A keby táto myseľ hľadala tieto zážitky *aj ich referenty*, bola by to dobrá a pravdivá budúcnosť; ale pretože sa vynechal tento *jeden rozmer* hodnoty, budúcnosť sa stala niečím tupým. Nudným a únavným, pretože aj keď táto myseľ *cíti*, že prežíva neuveriteľne nové zážitky, ani ten pocit nie je nijako pravdivý.

Alebo opačný problém – činiteľ, ktorý obsahuje všetky stránky ľudskej hodnoty, *okrem* cenenia si subjektívnej skúsenosti. Takže výsledkom je nevedomý optimalizátor, ktorý ide a robí skutočné objavy, ale tieto objavy si nikto nevychutnáva a neužíva, pretože nie je nikto, kto by to mohol robiť. Toto, priznávam, ani celkom neviem, či je možné. Vedomie ma stále do istej miery mätie. Ale vesmír, v ktorom nie je nikto, kto by ho vnímal, akoby ani nebol.

Hodnota nie je iba zložitá, ale aj *krehká*. Existuje *viac než jeden rozmer* ľudskej hodnoty, kde *ak stratíme iba túto jednu vec*, celá Budúcnosť zostane nulová. *Jediný* úder, a *celá* hodnota sa roztriešti. Nie každý *jeden* úder roztriešti *celú* hodnotu – ale existuje viac než jeden možný „jediný úder", ktorý to urobí.

A potom sú tu dlhé obrany týchto postojov, ktoré sa opierajú o 75 % mojich článkov na *Overcoming Bias*, takže by to bola práca na viac ako jeden deň všetky ich zhrnúť. Možno niektorý iný týždeň. Existuje tak veľa vetiev, na aké som videl, že sa tá diskusia delí.

Napokon – myseľ by *nemala* len tak ísť a mať ten istý zážitok znovu a znovu a znovu. Iste by sa žiadna superinteligencia nemohla tak hrubo pomýliť ohľadom toho, čo je správne urobiť?

Prečo by nejaká supermyseľ chcela niečo také vnútorne bezcenné ako pocit objavu bez nejakého skutočného objavu? Dokonca aj keby to bola jej funkcia úžitku, nevšimla by si jednoducho, že jej funkcia úžitku je zlá, a neprepísala by ju? Má predsa slobodnú vôľu, nie?

Určite aspoň *nuda* musí byť univerzálna hodnota. Vyvinula sa v ľuďoch, lebo má hodnotu, nie? Takže ľubovoľná myseľ, ktorá by s nami nemala spoločný odpor voči opakovaniu, by nemohla prosperovať vo vesmíre a bola by odstránená...

Ak poznáte rozdiel medzi inštrumentálnymi hodnotami a konečnými hodnotami, *a* poznáte hlúposť prirodzeného výberu, *a* rozumiete, ako sa táto hlúposť prejavuje v rozdieli medzi vykonávaním adaptácií a maximalizovaním spôsobilosti, *a* viete, že to premenilo inštrumentálne subciele rozmnožovania na nepodmienené emócie mimo pôvodného kontextu...

...*a* viete, ako funguje vyváženosť medzi skúmaním a využívaním v umelej inteligencii...

...potom možno dokážete uvidieť, že ľudská forma nudy, ktorá si vyžaduje stály prísun noviniek kvôli samotným novinkám, nie je všeobecné univerzálne pravidlo, ale iba konkrétny algoritmus, ktorý do nás vypľula evolúcia. A možno dokážete uvidieť, že prevažná väčšina možných maximalizátorov očakávaného úžitku by sa venovala iba obmedzenému množstvu efektívneho skúmania a trávila by väčšinu svojho času ťažením zo zatiaľ najlepšej nájdenej alternatívy, znovu a znovu a znovu.

To je ale dosť základných informácií.

A tak ďalej, a tak ďalej, a tak ďalej cez 75 % mojich článkov na *Overcoming Bias*, a mnoho reťazí klamov a protivysvetlení. Niektorý týždeň sa možno pokúsim nakresliť celý ten diagram. Ale momentálne budem predpokladať, že ste čítali tie argumenty, a iba predložím záver:

Nemôžeme uvoľniť naše držanie budúcnosti – pustiť kormidlo – tak, aby nám stále zostalo niečo hodnotné.

A tí, ktorí si myslia, že *môžu*...

...sa snažia byť kozmopolitní. Ja to chápem. Čítal som tie isté knihy vedeckej fantastiky ako dieťa: Provinční darebáci, ktorí zotročujú mimozemšťanov pre ten zločin, že nevyzerajú presne ako ľudia. Provinční darebáci, ktorí zotročujú bezmocné UI vo väzeniach, predpokladajúc, že kremík nemôže mať vedomie. A kozmopolitní hrdinovia, ktorí rozumejú, že *myseľ nemusí byť rovnaká ako my, aby sme ju prijali ako hodnotnú*...

Čítal som tie knihy. Kedysi som im veril. Ale krása, ktorá vyskočí z jednej krabičky, nevyskočí zo *všetkých* krabičiek. Ak zanecháte všetok poriadok, nezostane vám dokonalá odpoveď, ale dokonalý šum. Niekedy musíte zanechať staré pravidlo dizajnu, aby ste postavili lepšiu pascu na myši, ale to nie je to isté ako vzdať sa všetkých pravidiel dizajnu a zbierať drevené piliny na hromadu, lebo každý vzor dreva je rovnako dobrý ako hociktorý iný. Staré pravidlo sa vždy zanecháva na pokyn nejakého vyššieho pravidla, nejakého vyššieho vládnucej hodnoty.

Ak uvoľníte svoje držanie ľudskej morálky a metamorálky – výsledok nebude tajomný a mimozemský a krásny podľa štandardov ľudskej hodnoty. Je to morálny šum, vesmír vyplnený kancelárskymi spinkami. Aby ste sa zmenili z ľudskej morálky *smerom k zlepšeniu a nie k entropii*, potrebujete kritérium zlepšenia; a toto kritérium by bolo fyzicky reprezentované v našich mozgoch, a iba v našich mozgoch.

Uvoľnite držanie vesmíru ľudskou hodnotu, a skončí ako *vážne* bezcenný. Nie čudný a mimozemský a úžasný, šokujúci a desivý a krásny nad ľudskú predstavivosť. Jednoducho vyplnený kancelárskymi spinkami.

To totiž iba niektorí *ľudia*, ako vidíte, majú túto predstavu, že by sme mali prijať mnohoraké rozmanitosti mysle – že by sme mali chcieť, aby Budúcnosť bola niečo väčšie než minulosť – že by sme nemali byť pripútaní k svojmu starému ja – že by sme sa mali chcieť meniť a posúvať vpred.

Maximalizátor kancelárskych spiniek si iba vyberie tú činnosť, ktorá vedie k najväčšiemu množstvu kancelárskych spiniek.

Žiaden obed zadarmo. Chcete mať úžasný a tajomný vesmír? To je *vaša* hodnota. To je *vaša* práca, aby ste túto hodnotu vytvorili. Nechajte tú hodnotu pôsobiť silou cez vás, ktorí ju predstavujete, nechajte ju robiť vo vás rozhodnutia, ktoré vytvarujú budúcnosť. A možno naozaj získate úžasný a tajomný vesmír.

Žiaden obed zadarmo. Hodnotné veci sa objavujú preto, lebo systém cieľov, ktorí si ich cení, podnikol akciu, aby ich vytvoril. Kancelárske spinky sa maximalizátoru kancelárskych spiniek

nezhmotňujú z ničoho. A úžasne cudzia a tajomná Budúcnosť sa ľuďom nezhmotní z ničoho, ak naše hodnoty, ktoré juuprednostňujú, budú fyzicky zničené – alebo len *narušené* nesprávnym smerom. Potom vo vesmíre nezostane nič, čo by pracovalo na tom, aby ten vesmír urobilo hodnotnejším.

Vy *máte* hodnoty, dokonca aj keď sa snažíte byť „kozmopolitný“, pokúšate sa prejaviť správne cnostné ocenenie mimozemských myslí. Vaše hodnoty potom viac splývajú s neviditeľným pozadím – sú menej *očividne* ľudské. Váš mozog pravdepodobne ani nevygeneruje alternatívu takú strašnú, že by vás to prebudilo a donútilo povedať: „Nie! Niečo sa pokazilo!“ dokonca aj pri vašej maximálnej kozmopolitnosti. Napríklad: „nevedomý optimalizátor pohlcuje všetku hmotu vo svojom svetelnom kuželi a vypĺňa vesmír kancelárskymi spinkami.“ Vy si iba predstavujete zvláštne mimozemské svety, ktoré by ste si cenili.

Pokúšať sa byť „kozmopolitný“ - byť *občanom vesmíru* – iba odstraňuje *pozlátko na povrchu* z cieľov, ktoré vyzerajú *očividne* „ľudské“.

Ale ak nechcete, aby Budúcnosť bola vyplnená kancelárskymi spinkami, a radšej by ste mali civilizáciu plnú...

...vedomých bytostí...

...s príjemnými zážitkami...

...ktoré nie sú *tie isté* zážitky znovu a znovu dokola...

...a týkajú sa niečoho iného, okrem toho, že sú to iba postupnosti vnútorných príjemných pocitov...

...ktoré sa učia, objavujú, vyberajú si z možností...

...no, práve som presiel cez články o teórii zábavy, ktoré rozoberajú *niektoré* zo skrytých podrobností týchto krátkych anglických slov.

Hodnoty, ktoré by ste mohli chváliť ako *kozmopolitné* alebo *univerzálne* alebo *základné* alebo *samozrejmý zdravý rozum*, sú reprezentované vo vašom mozgu rovnako ako tie hodnoty, ktoré by ste zamietli ako *iba ľudské*. Tieto hodnoty pochádzajú z dlhej histórie ľudstva a z morálne zázračnej hlúposti evolúcie, ktorá nás vytvorila. (A keď som si toto *konečne* uvedomil, cítil som menšiu hanbu za hodnoty, ktoré vyzerali „provinčne“ - ale to je iná téma.)

Tieto hodnoty *nevznikajú* vo všetkých možných mysliach. *Neobjavia* sa len tak z ničoho, aby skritizovali a odvolali funkciu úžitku maximalizátora očakávaných kancelárskych spiniek.

Dotknite sa príliš silno v nesprávnom rozmere, a fyzické reprezentácie týchto hodnôt sa roztrieštia – a *nevrátia sa späť*, pretože nezostane nič, čo by ich *chcelo* priviesť späť.

A *referent* týchto hodnôt – hodnotný vesmír – už viac nebude mať žiaden fyzický dôvod vzniknúť.

Pustite kormidlo, a Budúcnosť havaruje.

*→

280. Dar, ktorý dáme zajtrajšku

Ako, ach ako, sa nemilujúcemu a nemysliacemu vesmíru podarilo vypľuť mysle, ktoré boli schopné lásky?

„To nie je žiadne tajomstvo,“ poviete, „je to iba vec prirodzeného výberu.“

Ale prirodzený výber je krutý, krvavý, a príšerne hlúpy. Ešte aj keď na povrchu vecí biologické organizmy *priamo* nebojujú jeden proti druhému – netrhajú jeden druhého *priamo* pazúrmi – stále je tam medzi génmi hlbšia súťaž. Genetická informácia sa vytvára, keď gény zvyšujú svoju *relatívnu* frekvenciu v nasledujúcej generácii - „genetickej spôsobilosti“ záleží nie na tom, koľko máte detí, ale či máte detí *viac* než ostatní. Je celkom možné, aby sa druh vyvinul k vyhynutiu, ak víťazné gény hrajú hry so záporným súčtom.

Ako, ach ako, mohol takýto proces vytvoriť bytosti schopné lásky?

„To nie je tajomstvo,“ poviete, „vo svete nikdy neexistujú tajomstvá; tajomnosť je vlastnosťou otázok, nie odpovedí. Deti majú s matkou spoločné gény, preto matka svoje deti miluje.“

→ http://lesswrong.com/lw/y3/value_is_fragile/

Ale niekedy matky adoptujú deti, a aj tak ich milujú. A matky milujú svoje deti kvôli nim samým, nie kvôli ich génom.

„To nie je tajomstvo," poviete, „Jednotlivé organizmy sú vykonávatelia adaptácií, nie maximalizátori spôsobilosti. Evolučná psychológia neznamená vedomé maximalizovanie spôsobilosti – počas väčšiny ľudských dejín sme ani nevedeli, že gény existujú. Nepočítali sme účinky našich činov na genetickú spôsobilosť vedome, dokonca ani podvedome."

Ale ľudské bytosti tvoria priateľstvá dokonca aj s nepríbuznými: ako, ach ako je to možné?

„To nie je tajomstvo, pretože lovci-zberači často hrali iterované Väzenské dilemy, na ktoré riešením je vzájomný altruizmus. Niekedy tým najnebezpečnejším človekom v celom kmeni nie je ten najsilnejší, najkrajší, dokonca ani najbystrejší, ale ten, kto má najviac spojencov."

Napriek tomu, nie všetci kamaráti sú kamarátmi iba do dobrého počasia; máme pojem skutočného priateľstva – a niektorí ľudia pre svojich priateľov obetovali život. Nemá azda takáto oddanosť sklon odstrániť sa z genofondu?

„Sám si to povedal: máme pojem skutočného priateľstva a kamarátstva do dobrého počasia. Vieme rozoznať, alebo skúšame rozoznať, rozdiel medzi niekým, kto nás považuje za vzácneho spojenca, a niekým, kto vykonáva adaptáciu kamarátstva. Nemáme skutočné priateľstvo s niekým, o kom si nemyslíme, že je skutočným priateľom voči nám – a niekto, kto má veľa *skutočných* priateľov, je hrozivejší než niekto, kto má veľa kamarátov do pekného počasia."

A Mohandas Gándhí, ktorý naozaj nastavil druhé líce? Tí, ktorí sa snažia slúžiť celému ľudstvu, či už im ľudstvo na oplátku poslúži alebo nie?

„To je možno omnoho zložitejší príbeh. Ľudia nie sú iba spoločenské živočíchy. Sme aj politické živočíchy, ktoré pomocou jazyka debatujú o pravidlách v adaptívnych kontextoch kmeňa. Niekedy ten najhrozivejší človek nie je ten najsilnejší, ale ten, kto vie najšikovnejšie argumentovať, že pravidlá, ktoré chce on, zodpovedajú tomu, čo chcú ostatní."

Ehm... toto nevysvetľuje Gándhího, alebo mi niečo uniklo?

„Pointa je, že máme schopnosť *argumentovať* o tom: ‚Čo by sa malo urobiť?' ako o *návrhu* – dokážeme tvoriť takéto argumenty a reagovať na takéto argumenty, a bez toho by nemohla existovať politika."

Dobre, ale čo Gándhí?

„Veril istým komplikovaným návrhom ohľadom: ‚Čo by sa malo urobiť?' a urobil to."

Toto znie, že by sa tým dalo vysvetliť hocijaké ľudské správanie.

„Keby sme sledovali naspäť reťaz kauzality cez všetky argumenty, obsahovala by: morálnu architektúru, ktorá má schopnosť argumentovať za *všeobecné abstraktné* morálne návrhy ako: ‚Čo by sa malo urobiť ľuďom?', odvolávať sa na zabudované intuície ako je spravodlivosť, pojem povinnosti, vyhýbanie sa bolesti + súcit; niečo ako uprednostňovanie jednoduchých morálnych návrhov, pravdepodobne ako znovupoužitie nášho Occamovského východiska; a konečným výsledkom tohto všetkého, plus azda aj efektov memetického výberu, bolo: ‚Nemal by si ubližovať ľuďom' v úplnej všeobecnosti..."

A takto vznikol Gándhí.

„Pokiaľ si nemyslíš, že to bolo kúzlo, musí to nejako zapadať do zákonitého kauzálneho vývoja vesmíru."

No... iste by som nepredpokladal kúzlo, pod nijakým názvom.

„Dobre."

Ale predsa... nevyzerá to trochu... *úžasne*... že stovky miliónov rokov turnajov smrti evolúcie dokázali vypľuť matky a otcov, sestry a bratov, manželov a manželky, spoľahlivých priateľov a čestných nepriateľov, skutočných altruistov a strážcov káuz, dokonca umelcov obetujúcich sa pre svoje umenie, všetkých praktizujúcich tak veľa druhov lásky? Voči toľkým iným veciam než gény? Robia svoju časť, aby bol ich svet menej hnusný, niečo iné okrem mora krvi a násilia a nezmyselného replikovania?

„Tvrdíš, že ťa to prekvapuje? Ak áno, zamysli sa nad svojím základným modelom, pretože ťa doviedol k prekvapeniu zo skutočného stavu vecí. Od samotného začiatku sa nestala jediná nezvyčajná vec.“

Ale ako by toto *nebolo* prekvapivé?

„Naznačuješ azda, že nejaká postava stála v tieni za kulisami a riadila evolúciu?“

Sakra nie. Ale...

„Pretože *keby* si to naznačoval, musel by som sa opýtať, ako sa táto postava v tieni *pôvodne* rozhodla, že láska je *žiadúcim* výsledkom evolúcie. Musel by som sa opýtať, kde táto postava dostala preferencie, ktoré zahŕňajú veci ako láska, priateľstvo, vernosť, spravodlivosť, česť, romantika, a tak ďalej. V evolučnej psychológii môžeme vidieť, ako vznikli *tieto konkrétne výsledky* – ako boli *v prvom rade vytvorené tieto konkrétne ciele a nie iné*. Môžeš to nazývať ‚prekvapujúce‘, ak chceš. Ale ak naozaj rozumieš evolučnej psychológii, môžeš vidieť, ako rodičovská láska a romantika a česť, a dokonca skutočný altruizmus a morálne argumenty *nesú konkrétnu pečať dizajnu prirodzeného výberu* v konkrétnych adaptívnych kontextoch lovcov-zberačov zo savany. Ak teda bola nejaká postava v tieni, ona sama sa musela vyvinúť – a to odstraňuje celý dôvod, prečo ju predpokladať.“

Ja nepredpokladám postavu v tieni! Iba sa pýtam, ako mohli ľudia skončiť takto *pekne*.

„Pekne! Pozeral si sa v poslednom čase na túto planétu? Máme aj všetky tie ostatné emócie, ktoré sa tiež vyvinuli – ktoré by ti povedali veľmi jasne, že sme sa vyvinuli, keby si o tom začal pochybovať. Ľudia nie sú vždy milí.“

Sme čertovsky milší než ten proces, ktorý nás vytvoril, ktorý necháva slony vyhladovať na smrť, keď im vypadajú zuby, a nezbaví bolesti gazelu ani keď leží a umiera a teda je už evolúcii tak či onak nepotrebná. Netreba veľa na to byť milší než evolúcia. Mať *teoretickú schopnosť* urobiť jediné gesto milosrdenstva, cítiť jediný záchvev súcitu, už znamená byť milší než evolúcia.

Ako mohla evolúcia, ktorá je sama taká necitlivá, vytvoriť mysle na tak kvalitatívne vyššej morálnej úrovni než je ona sama? Ako mohla evolúcia, ktorá je taká hnusná, skončiť urobením niečoho tak *krásneho*?

„Krásne, hovoríš? Bachova *Malá fúga v G mol* môže byť krásna, ale zvukové vlny, ako cestujú vzduchom, nie sú označené malými značkami, ktoré by udávali ich krásu. Ak chceš nájsť *výslovne zakódovanú* správu o tom, nakoľko je táto fúga krásna, musíš sa pozrieť na ľudský mozog – nikde inde vo vesmíre ju nenájdeš. Takýto úsudok nenájdeš ani v moriach ani na horách: nie sú to mysle, nevedia myslieť.“

Azda je to tak. Napriek tomu nám evolúcia *naozaj dala* schopnosť obdivovať krásu kvetov. To stále vyzerá, že si vyžaduje hlbšiu odpoveď.

„Nevidíš kruhovosť svojej otázky? Keby krása bola nejaké veľké svetlo na oblohe, ktoré žiari odinakiaľ než z ľudí, potom by tvoja otázka mohla dávať zmysel – hoci by stále bola otázka, ako ľudia začali vnímať toto svetlo. Vyvinuli ste sa s psychológiou inou ako je evolúcia: Evolúcia nemá nič také ako je inteligencia alebo presnosť potrebná na presné zopakovanie svojho systému cieľov. Pri vypľutí prvej skutočnej mysle sa jednoduché kritérium spôsobilosti evolúcie rozbilo na tisíc hodnôt. Vyvinuli ste sa s psychológiou, ktorá pripisuje úžitok veciam, o ktoré sa evolúcia nestará, ako je ľudský život a šťastie. A potom sa obzriete a poviete: ‚Aké podivuhodné, že necitlivá evolúcia vytvorila mysle, ktorým záleží na cítiacom živote!‘ Takže vaše veľké čudovanie a žasnutie, ktoré vyzerá ako príliš veľká náhoda, v skutočnosti nie je žiadna náhoda.“

Ale aj tak je stále úžasné, že na svete vznikla práve táto konkrétna slučka, a nie nejaká iná. Že oslavujeme veci ako láska a nie nenávisť, krása a nie škaredosť.

„Myslím, že ma tu nepočúvaš. To vám pripadá prirodzené uprednostňovať krásu a altruizmus ako špeciálne, ako žiadúce, pretože si ich vysoko ceníte; a nevnímate to ako nezvyčajný fakt o vás samotných, lebo rovnako to má veľa vašich priateľov. Preto očakávate, že aj duch dokonalej prázdnoty by si cenil život a šťastie – a potom, z tohto pohľadu mimo skutočnosti, by sa naozaj bola stala veľká náhoda.“

Ale veď môžeš vytvoriť argumenty o dôležitosti krásy a altruizmu z prvotných princípov – že nás náš zmysel pre estetiku vedie k vytváraniu novej zložitosti, namiesto opakovania tých istých vecí dokola

a dokola; a že altruizmus je dôležitý, lebo nás vyvádza zo seba samých, dáva nášmu životu vyšší zmysel než je číre zvieracie sebectvo.

„Ach, a *tento* argument by pohol dokonca aj duchom dokonalej prázdnoty – lebo už si sa odvolal na trochu odlišné hodnoty? To nie sú prvotné princípy, sú to len *iné* princípy. Hoci si prevzal pompézny filozofický tón, stále neexistujú žiadne *všeobecne* presvedčivé argumenty. Jediné, čo si urobil, bolo odovzdanie rekurzívnej mince."

Nemyslíš si, že sme sa nejako vyvinuli, aby sme *dosiahli* niečo, čo je mimo...

„Načo je dobré predpokladať niečo mimo? Prečo by sme mali tejto veci mimo venovať viac pozornosti než našej existencii ako ľudí? Ako mení tvoju osobnú zodpovednosť, keď povieš, že si iba poslúchal príkazy nejakej veci mimo? A stále by si sa musel vyvinúť, aby si nechal túto vec mimo, a nie niečo iné, ovládať tvoje činy. Bola by to *príliš veľká náhoda.*"

Príliš veľká náhoda?

„Kvet je krásny, hovoríš. Myslíš si, že za touto krásou nie je žiaden príbeh, alebo že veda tento príbeh nepozná? Peľ z kvetov prenášajú včely, takže sa kvety sexuálnym výberom vyvinuli tak, aby priťahovali včely – zhodou okolností napodobňovaním niektorých rozmnožovacích signálov včiel; vzory kvetov by vyzerali omnoho zložitejšie, keby si videl ultrafialovú. Zdravé kvety sú signálom úrodnej zeme, v ktorej pravdepodobne rastie ovocie a iné poklady, a pravdepodobne je tam aj korisť; je teda také čudné, že sa ľudia vyvinuli, aby sa im páčili kvety? Ale aby existovalo nejaké veľké svetlo napísané v samotných hviezdach – tých veľkých nemysliacich guliach horiaceho vodíka – ktoré by *tiež* hovorilo, že kvety sú krásne, *to* už by bola príliš veľká náhoda."

Takže si vyvrátil krásu kvetov?

„Nie, vysvetlil som ju. Samozrejme, že existuje príbeh za krásou kvetov a faktom, že ich považujeme za krásne. Za usporiadanými udalosťami sa nachádzajú usporiadané príbehy; a čo nemá žiaden príbeh, to je produktom náhodného šumu, čo nie je o nič lepšie. Ak sa nedokážeš tešiť z vecí, ktoré majú za sebou príbeh, potom bude tvoj život prázdny. Nemyslím si, že mám z kvetov menšiu radosť než ty; možno väčšiu, pretože sa teším aj z toho príbehu."

Možno, ako hovoríš, to nie je prekvapenie z kauzálneho hľadiska – narušenie fyzikálneho poriadku vesmíru. Ale stále mi pripadá, že pri tomto stvorení ľudí evolúciou sa stalo niečo, čo je vzácne a obdivuhodné a úžasné. Ak to nemôžeme označiť za fyzikálny zázrak, označme to za morálny zázrak.

„Pretože je to zázrak iba z pohľadu morálky, ktorá tak vznikla, čím sa vyvrátili všetky tie domnelé náhody z čisto kauzálneho a fyzikálneho pohľadu?"

No... asi by si tie slová mohol vysvetliť takto, áno. Myslel som tým len, že niečo je mimoriadne prekvapujúce a úžasné na úrovni morálky, hoci to nie je prekvapujúce na úrovni fyziky.

„Myslím, že to som povedal."

Ale aj tak sa mi zdá, že zo svojho vlastného pohľadu, niečo z toho úžasu odvádzaš preč.

„Potom máš problém tešiť sa z púhej skutočnosti. Láska musela nejako začať. Musela niekde vstúpiť do vesmíru. Je to ako pýtať sa, ako vznikol samotný život – a hoci ty si narodil zo svojho otca a matky, a oni zase vznikli zo svojich živých rodičov, ak sa vrátiš ďaleko a ďaleko a ďaleko naspäť, nakoniec prídeš k replikátoru, ktorý vznikol čírou náhodou – k hranici medzi životom a neživotom. Tak je to aj s láskou."

„Zložitý vzor musí byť vysvetlená príčinou, ktorá ešte nie je týmto zložitým vzorom. Musí byť vysvetlená nielen tá udalosť, ale aj jej skutočný tvar a podoba. Aby láska po prvýkrát vstúpila do Času, musela prísť z niečoho, čo nie je láska; keby toto nebolo možné, potom by láska nemohla existovať."

„Rovnako ako život samotný vyžadoval, aby prvý replikátor vznikol náhodou, bez rodičov, ale aj tak zapríčinený: ďaleko, ďaleko späť v kauzálnej reťazi, ktorá viedla k tebe: pred 3,85 miliardami rokov, v nejakom malom prílivovom jazierku."

„Možno sa deti vašich detí budú pýtať, ako je možné, že sú schopné lásky."

„A ich rodičia povedia: Pretože my, ktorí milujeme, sme vás stvorili, aby ste milovali."

„A deti vašich detí sa budú pýtať: Ale ako je možné, že *vy* milujete?"

„A ich rodičia odpovedia: Pretože naši vlastní rodičia, ktorí tiež milovali, nás stvorili, aby sme tiež milovali.“

„Potom sa deti vašich detí opýtajú: Ale kde sa to celé začalo? Kde končí táto rekurzia?“

„A ich rodičia povedia: Kde bolo, tam bolo, veľmi dávno a veľmi ďaleko, pred dávnymi vekmi, existovali inteligentné bytosti, ktoré samotné neboli inteligentne nadizajnované. Kde bolo, tam bolo, existovali milujúce tvory vytvorené niečím, čo nemilovalo.“

„Kde bolo, tam bolo, keď celá civilizácia bola v jedinej galaxii a na jedinej hviezde: a na jedinej planéte, na mieste zvanom Zem.“

„Dávno a ďaleko, pred dávnymi vekmi.“

*→

→ http://lesswrong.com/lw/sa/the_gift_we_give_to_tomorrow/

W: Kvantifikovaný humanizmus

281. Necitlivosť k rozsahu

Jedného dňa sa vedci pýtali troch skupín pokusných osôb, koľko by boli ochotné zaplatiť za záchranu 2 000 / 20 000 / 200 000 sťahovavých vtákov pred utopením v neprikrytých ropných nádržiach. Jednotlivé skupiny odpovedali: 80 dolárov, 78 dolárov, a 88 dolárov.[250] Toto je *necitlivosť k rozsahu*, alebo *zanedbanie rozsahu*: počet zachránených vtákov – *rozsah* altruistického činu – málo vplýva na ochotu platiť.

Podobné experimenty ukázali, že obyvatelia Toronta by za vyčistenie všetkých znečistených jazier v Ontariu zaplatili len o málo viac než za vyčistenie znečistených jazier v jednom okrese Ontaria,[251] alebo že obyvatelia štyroch západných štátov USA by zaplatili iba o 28% viac za ochranu všetkých 57 prírodných rezervácií než za ochranu jedinej rezervácie.[252] Ľudia si predstavia „jedného vyčerpaného vtáka s perami nasiaknutými čiernou ropou, neschopného uniknúť".[253] Tento obraz či *prototyp* vyvolá určitý stupen emocionálneho vzrušenia, ktoré je v prvom rade zodpovedné za ochotu platiť – a tento obraz je vo všetkých prípadoch rovnaký. Čo sa rozsahu týka, ten vyhadzujeme von oknom – žiaden človek si nevie naraz predstaviť 2 000 vtákov, tobôž 200 000. Zvyčajné zistenie je, že *exponenciálne* zvýšenie rozsahu vedie k *lineárnemu* zvýšeniu ochoty platiť – čo možno zodpovedá lineárnemu zvýšeniu času, ktorý našim očiam trvá preletieť všetky tie nuly; toto malé úsilie sa k afektu z prototypu pripočíta, namiesto aby sa ním vynásobil. Táto hypotéza sa nazýva „ohodnocovanie podľa prototypu".

Alternatívna hypotéza je „nákup morálneho uspokojenia". Ľudia minú dosť peňazí na to, aby si kúpili *hrejivý pocit*, že splnili svoju povinnosť. Úroveň nákladov potrebných na dosiahnutie hrejivého pocitu závisí od osobnosti a finančnej situácie, ale určite nemá nič spoločné s počtom vtákov.

Voči rozsahu sme necitliví aj keď ide o ľudské životy: Zvýšenie domnelého rizika z pitia chlorovanej vody z 0,004 na 2,43 smrtí ročne z 1000 – na 600-násobok – zvýšilo ochotu platiť z 3,78 dolára na 15,23.[254] Baron a Greene nenašli žiaden účinok pri zmene počtu zachránených životov na 10-násobok.[255]

Článok s názvom „Necitlivosť voči hodnote ľudského života: Štúdia psychofyzického otupenia" zozbieral dôkazy, že naše vnímanie ľudskej smrti sa riadi Weberovým zákonom – nasleduje logaritmickú škálu, kde „najmenší vnímateľný rozdiel" je konštantným zlomkom celku. Navrhovaný zdravotný program na záchranu životov utečencov z Rwandy si získal omnoho vyššiu podporu, keď sľuboval zachrániť 4 500 životov z tábora 11 000 utečencov, než 4 500 životov z tábore 250 000 utečencov. Potenciálny liek na chorobu musel sľubovať zachrániť omnoho viac životov, aby bol považovaný za hoden financovania, ak sa povedalo, že choroba už zabíla 290 000 namiesto 160 000 alebo 15 000 ľudí ročne.[256]

---

250 William H. Desvousges et al., *Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy*, technical report (Research Triangle Park, NC: RTI International, 2010), doi:10.3768/rtipress.2009.bk.0001.1009.

251 Daniel Kahneman, „Comments by Professor Daniel Kahneman," in *Valuing Environmental Goods: An Assessment of the Contingent Valuation Method*, ed. Ronald G. Cummings, David S. Brookshire, and William D. Schulze, vol. 1.B, Experimental Methods for Assessing Environmental Benefits (Totowa, NJ: Rowman & Allanheld, 1986), 226–235, http://yosemite.epa.gov/ee/epa/eerm.nsf/vwAN/EE-0280B-04.pdf/$file/EE-0280B-04.pdf.

252 Daniel L. McFadden and Gregory K. Leonard, „Issues in the Contingent Valuation of Environmental Goods: Methodologies for Data Collection and Analysis," in *Contingent Valuation: A Critical Assessment*, ed. Jerry A. Hausman, Contributions to Economic Analysis 220 (New York: North-Holland, 1993), 165–215, doi:10.1108/S0573-8555(1993)0000220007.

253 Kahneman, Ritov, and Schkade, „Economic Preferences or Attitude Expressions?"

254 Richard T. Carson and Robert Cameron Mitchell, „Sequencing and Nesting in Contingent Valuation Surveys," *Journal of Environmental Economics and Management* 28, no. 2 (1995): 155–173, doi:10.1006/jeem.1995.1011.

255 Jonathan Baron and Joshua D. Greene, „Determinants of Insensitivity to Quantity in Valuation of Public Goods: Contribution, Warm Glow, Budget Constraints, Availability, and Prominence," *Journal of Experimental Psychology: Applied* 2, no. 2 (1996): 107–125, doi:10.1037/1076-898X.2.2.107.

Ponaučenie: Ak chcete byť efektívnym altruistom, musíte si to premyslieť tou časťou mozgu, ktorá spracováva tie nudné atramentové nuly na papieri, nie iba tou časťou, ktorá sa veľmi znepokojuje ohľadom úbohého trápiaceho sa vtáka zmáčaného ropou.

*→

282. Jeden život proti celému svetu

„Ktokoľvek zachráni jediný život, je to akoby zachránil celý svet."

--Talmud, Sanhedrin 4:5

Krásna myšlienka, nie? Vychutnajte si ten hrejivý pocit.

Môžem dosvedčiť, že pomôcť jednému človeku vám dá rovnako dobrý *pocit* ako pomôcť celému svetu. Jedného dňa, keď som bol celý deň vyhorený a zabíjal som čas na internete – je to trochu komplikované, ale v zásade som niekomu pomohol zmeniť celý život zanechaním anonymného komentára na blogu. Nečakal som, že to bude mať takýto veľký účinok, ale stalo sa. Keď som zistil, čo som dosiahol, dalo mi to *obrovsky* povznášajúci pocit. Eufória trvala celý deň až do noci; až ďalšie ráno začala vyprchávať. Bolo to rovnako príjemné (a toto je tá strašidelná časť) ako eufória z významného vedeckého objavu, čo bola dovtedy moja najlepšia predstava o tom, ako sa asi môže cítiť človek na drogách.

Zachrániť niekomu život pravdepodobne *je* rovnako dobrý pocit ako byť prvým človekom, ktorý si uvedomí, prečo hviezdy svietia. Pravdepodobne to *je* rovnako dobrý pocit ako zachrániť celý svet.

Avšak, milý čitateľ, ak budeš mať niekedy na výber medzi záchranou jediného života a záchranou celého sveta – potom zachráň celý svet. Prosím. Pretože za týmto hrejivým pocitom je *hovadsky gigantický rozdiel*. Pre niektorých ľudí je predstava, že zachrániť celý svet je *výrazne lepšie* než zachrániť jeden ľudský život samozrejmá, ako povedať, že šesť miliárd dolárov má väčšiu hodnotu ako jeden dolár, alebo že šesť kubických kilometrov zlata váži viac ako jeden kubický meter zlata. (A to nerátame očakávanú hodnotu potomstva zachránených.) Prečo by to *nemalo* byť samozrejmé? Nuž, predpokladajme, že existuje kvalitatívna povinnosť zachrániť toľko ľudských životov, koľko môžete – potom ten, kto zachráni celý svet a ten, kto zachráni jeden ľudský život, si iba plnia tú istú povinnosť. Alebo predpokladajme, že sa namiesto konsekvencializmu riadime gréckym poňatím osobnej cnosti; ten, kto zachráni celý svet je cnostný, ale nie šesť miliárdkrát cnostnejší než ten, kto zachráni jeden ľudský život. Alebo povedzme, že už hodnota jedného ľudského života je priveľká nad naše chápanie – taká, že pominuteľný žiaľ, ktorý prežívame na pohreboch, je nekonečným podcenením toho, čo sa stratilo – a teda žialiť nad celým svetom mení len málo.

Súhlasím, že ľudský život má nepredstaviteľne vysokú hodnotu. Tvrdím aj, že dva ľudské životy sú dvakrát tak nepredstaviteľne hodnotné. Alebo inými slovami: Ktokoľvek zachráni jeden život, je to akoby zachránil celý svet; ktokoľvek zachráni desať životov, je to akoby zachránil desať svetov. Ktokoľvek *naozaj* zachráni celý svet – aby sme si to nemýlili s rétorickým akože zachraňovaním sveta – ten akoby zachránil intergalaktickú civilizáciu.

Dve hluché deti spia na železničných koľajach, vlak sa približuje; vidíte to, ale ste príliš ďaleko a nemôžete tie deti zachrániť. Ja som nablízku, preto vyskočím a odtiahnem jedno dieťa z koľají... a potom sa zastavím a pokojne si popíjam diétnu Pepsi, ako sa vlak rúti na to druhé dieťa. „*Rýchlo!*" kričíte na mňa. „*Urob niečo!*" Lenže (zavolám na vás) ja už som zachránil jedno dieťa z koľají a preto mám „nepredstaviteľne" veľa dobrých bodov. Či už zachránim to druhé dieťa alebo nie, stále budem konateľom „nepredstaviteľne" dobrého skutku. Nemám už teda žiaden motív konať. Neznie to správne, však?

256 David Fetherstonhaugh et al., „Insensitivity to the Value of Human Life: A Study of Psychophysical Numbing," *Journal of Risk and Uncertainty* 14, no. 3 (1997): 283–300, doi:10.1023/A:1007744326393.

→ http://lesswrong.com/lw/hw/scope_insensitivity/

Prečo by malo byť iné, ak filantrop venuje 10 miliónov dolárov na vyliečenie zriedkavej, ale okázalo smrteľnej choroby, ktorá postihuje iba sto ľudí na celej planéte, keď tie isté peniaze mohli s rovnakou pravdepodobnosťou vytvoriť liek pre menej okázalú chorobu, ktorá zabíja 10% zo 100 000 ľudí? Nemyslím si, že to je iné. Keď sa jedná o ľudské životy, máme povinnosť *maximalizovať*, nie uspokojovať; a táto povinnosť má rovnakú silu ako pôvodná povinnosť zachraňovať životy. Ktokoľvek sa vedome rozhodne zachrániť jeden život, keď mohol zachrániť dva – a nehovorím o tisíc životoch, alebo celom svete – zatratil sa rovnako ako hocijaký vrah.

Nie je kognitívne jednoduché míňať peniaze na záchranu životov, pretože otrepané metódy, ktoré vám okamžite prídu na um, nefungujú alebo škodia. (Neskôr napíšem, prečo to zvykne byť takto.) Stuart Armstrong tiež poukazuje, že ak ideme pohŕdať filantropom, ktorý neefektívne minie svoje životné úspory, mali by sme merať rovnakým metrom a pohŕdať ešte viac tými, ktorí mohli minúť peniaze na záchranu ľudských životov, ale neurobili to.

*→

283. Allaisov paradox

Vyberte si medzi nasledujúcimi dvoma možnosťami:

1A. Výhra 24 000 dolárov, naisto.

1B. Šanca 33/34 vyhrať 27 000 dolárov, a šanca 1/34 nevyhrať nič.

Ktorá z nich vás intuitívne viac priťahuje? A ktorú by ste si vybrali v skutočnom živote? A teraz, ktorá z týchto dvoch možností vás intuitívne viac priťahuje, a ktorú by ste si vybrali v skutočnom živote?

2A. Šanca 34% vyhrať 24 000 dolárov, a šanca 66% nevyhrať nič.

2B. Šanca 33% vyhrať 27 000 dolárov, a šanca 67% nevyhrať nič.

Allaisov paradox – ako ho Allais nazval, hoci to nie je naozaj paradox – bol jedným z prvých experimentálne odhalených konfliktov medzi teóriou rozhodovania a ľudským rozmýšľaním, v roku 1953.[257] Jemne som ho upravil, aby sa ľahšie matematicky počítal, ale základný problém je ten istý: Väčšina ľudí dáva prednosť 1A pred 1B, a väčšina ľudí dáva prednosť 2B pred 2A. Dokonca aj keď dáte obe otázky jednej pokusnej osobe, väčšina pokusných osôb vyjadrí obe preferencie zároveň.

Toto je problém, pretože možnosti číslo 2 sú rovné tretinovej šance hrať číslo 1. Čiže 2A je ekvivalentné pravdepodobnosti 34% hrať 1A, a 2B je ekvivalentné pravdepodobnosti 34% hrať 1B.

Medzi axiómami používanými na dôkaz, že „konzistentných“ rozhodujúcich sa možno vnímať ako maximalizujúcich očakávaný úžitok, je axióma nezávislosti: Ak je X lepšie ako Y, potom pravdepodobnosť P udalosti X a pravdepodobnosť (1 – P) udalosti Z musí byť lepšia ako pravdepodobnosť P udalosti Y a pravdepodobnosť (1 – P) udalosti Z.

Všetky tieto axiómy sú dôsledkami, ako aj podmienkami, konzistentnej funkcie úžitku. Musí byť teda možné dokázať, že pokusné osoby v uvedenom príklade *nemôžu* mať konzistentnú funkciu úžitku pre výsledky. Veru, nemôžete mať zároveň:

$$U(24\,000\ \$) > 33/34\ U(27\,000\ \$) + 1/34\ U(0\ \$)$$

$$0.34\ U(24\,000\ \$) + 0.66\ U(0\ \$) < 0.33\ U(27\,000\ \$) + 0.67\ U(0\ \$)$$

Tieto dve rovnice sú algebraicky nekonzistentné, bez ohľadu na U, takže Allaisov paradox nemá nič spoločné s klesajúcim hraničným úžitkom peňazí.

Maurice Allais pôvodne obhajoval prejavené preferencie poksuných osôb – vnímal tento pokus ako odhalenie chybu v konvenčných predstavách o úžitku, namiesto odhalenia chyby v ľudskej psychológii.

→ http://lesswrong.com/lw/hx/one_life_against_the_world/

257 Maurice Allais, „Le Comportement de l’Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l’Ecole Americaine,“ *Econometrica* 21, no. 4 (1953): 2, doi:10.2307/1907921; Daniel Kahneman and Amos Tversky, „Prospect Theory: An Analysis of Decision Under Risk,“ *Econometrica* 47 (1979): 263–292.

Predsa len to bolo v roku 1953, keď sa hnutie heuristík a skreslení nerozbehlo ešte ďalšie dve desaťročia. Allais si myslel, že tento pokus skrátka ukázal, že axióma nezávislosti očividne nie je dobrý nápad v skutočnom živote.

(Aká naivná, aká hlúpa, aká primitívna je bayesovská teória rozhodovania...)

Nepochybne by *istota* vlastnenia 24 000 dolárov mala *niečo* zavážiť. *Cítite* ten rozdiel, však? To pevné ubezpečenie?

(Začínam toto vnímať ako „naivný filozofický realizmus“ - predpoklad, že naše intuície priamo odhaľujú pravdy o tom, ktoré stratégie sú múdrejšie, akoby to bol priamo vnímaný fakt, že „1A je lepšie ako 1B“. Intuície *priamo* odhaľujú pravdy o ľudských kognitívnych funkciách a iba *nepriamo* odhaľujú (potom, čo sa zamyslíme nad kognitívnymi funkciami samotnými) pravdy o rozumnosti.)

„Ale počuj,“ poviete, „je to naozaj také hrozné odchýliť sa od bayesovskej krásy?“ Dobre, tak sa tie pokusné osoby neriadili našou peknou „axiómou nezávislosti“, ktorú hlásajú ľudia ako Neumann a Morgenstern. Kto však tvrdí, že veci *musia* byť pekné a upratané?

Načo sa ondiť s eleganciou, keď nás núti riskovať, aj keď nechceme? Očakávaný úžitok nám hovorí, že by sme mali výsledku priradiť nejaké číslo, potom toto číslo vynásobiť pravdepodobnosťou výsledku, sčítať to, atď. Dobre, ale prečo by sme to *mali* robiť? Prečo si radšej namiesto toho nevymyslieť nejaké stráviteľnejšie pravidlá?

Za opustenie bayesovskej cesty sa vždy platí nejaká cena. O tomto sú všetky tie vety o korehencii a jedinečnosti.

V tomto prípade, ak činiteľ uprednostňuje 1A pred 1B a 2B pred 2A, dostávame nejakú formu *obrátenia preferencií – dynamickú nekonzistenciu* v plánovaní tohto činiteľa. Stávate sa *pumpou na peniaze*.

Predstavme si, že o 12:00 hodím stostennou kockou. Ak na kocke padne číslo väčšie ako 34, hra končí. V opačnom prípade o 12:05 pozriem na prepínač, ktorý má dve nastavenia, A a B. Ak je nastavenie A, vyplatím vám 24 000 dolárov. Ak je nastavenie B, hodím 34-stennou kockou a vyplatím vám 27 000 dolárov, pokiaľ na kocke nepadlo „34“, lebo v tom prípade vám nevyplatím nič.

Predpokladajme, že dávate prednosť 1A pred 1B, a 2B pred 2A, a že by ste boli ochotní za tieto preferencie zaplatiť aspoň cent. Prepínač je za načiatku v polohe A. Pred 12:00 by ste mi zaplatili cent za to, aby som prepínač prepol na B. Na kocke padne 12. Medzi 12:00 a 12:05 by ste mi zaplatili cent za to, aby som prepol prepínač na A.

Vaše názory vás stáli dva centy.

Ak sa naozaj budete riadiť svojimi intuíciami, a odmietnete púhu eleganciu ako nezmyselnú posadnutosť poriadkom, potom nebuďte prekvapení, keď vám začnú unikať centy...

(Myslím si, že rovnaké zlyhanie pri proporcionálnom zmenšovaní emocionálneho dopadu malých pravdepodobností môže aj za lotérie.)

*→

284. Zase Allais!

Ha! Nečakal som takúto reakciu. Zdá sa, že som narazil na inferenčnú vzdialenosť.

Pri interpretácii Allaisovho paradoxu asi pomáha osvojiť si viac z *postoja* oblasti heuristík a skreslení, ako napríklad:

- Pokusné osoby majú sklon obhajovať nekoherentné preferencie, aj keď sú *naozaj* hlúpe.
- Ľudia prisudzujú veľkú hodnotu zmenám pravdepodobnosti od 0 alebo 1 (efekt istoty).

Začnime s témou nekoherentných preferencií – obrátenia preferencií, dynamická nekonzistencia, pumpy na peniaze, takéto veci.

Každý, kto vie niečo o teórii vyhliadky, nebude mať problém vytvoriť prípady, kde ľudia povedia, že by radšej hrali lotériu A ako lotériu B; ale keď sa ich opýtate, koľko by za to zaplatili, dajú viac za

→ http://lesswrong.com/lw/my/the_allais_paradox/

lotériu B ako za lotériu A. Rôzne časti vnemu sa zvýrazňujú, keď sa pýtate: „Čo by ste radšej?“ v priamom porovnaní a „Koľko by ste zaplatili?“ za jednu položku.

Teraz sa sťahujem, takže mám zbalené knihy, ale pokiaľ si pamätám, toto by malo zvyčajne vytvoriť prevrátenie preferencií:

1. 1/3 vyhrať 16 dolárov a 2/3 prehrať 2 doláre
2. 99/100 vyhrať 4 doláre a 1/100 prehrať 1 dolár

Pokiaľ si spomínam, väčšina ľudí by radšej hrala 2 ako 1. Ale ak ich požiadate o samostatné ohodnotenie stávok – požiadate ich o cenu, pri ktorej by im bolo jedno, či dostanú tie peniaze alebo dostanú šancu hrať túto hru – ľudia uvedú vyššiu cenu pre 1 než pre 2.[258]

Najprv im teda predáte šancu hrať stávku 1, za ich určenú cenu. Potom im ponúknete vymeniť stávku 1 za stávku 2. Potom od nich naspäť odkúpite stávku 2, za ich určenú cenu. Potom to urobíte znovu. Odtiaľ názov „pumpa na peniaze“.

Alebo parafrázujúc Stevea Omohundra: Ak by ste boli radšej v Oaklande než v San Franciscu, a radšej by ste boli v San Jose než v Oklande, a radšej by ste boli v San Franciscu než v San Jose, potom miniete hrozne veľa peňazí na taxík.

Čo je úžasné, ľudia *obhajujú* tieto vzorce preferencií. Niektoré pokusné osoby sa ich vzdajú, keď sa im ukáže efekt pumpovania peňazí – upravia svoje ceny alebo svoje preferencie – ale niektoré pokusné osoby ich obhajujú.

V jednom prípade gambleri v Las Vegas hrali takéto stávky o skutočné peniaze, používajúc ruletu. Potom sa im jeden z výskumníkov pokúsil vysvetliť problém s nekoherenciou medzi ich cenami a ich voľbami. Zo záznamu:[259,260]

> *Sarah Lichtenstein:* „Takže čo so stávkou A? Aké z nej máte pocity teraz, keď viete, že ste si vyberali jedno, ale ponúkali ste viac za druhé?“
>
> *Pokusná osoba:* „Je to čudné, ale vôbec mi to nevadí. Je to proste tak. Ukazuje to, že môj proces rozmýšľania nie je taký dobrý, ale inak... nevadí.“
>
> …
>
> *Lichtenstein:* „Mohla by som vás presvedčiť, že je to nerozumný vzorec?“
>
> *Pokusná osoba:* „Nie, nemyslím si, že by sa vám to podarilo, ale môžete skúsiť.“
>
> …
>
> *Lichtenstein:* „Navrhnem vám teda niečo, čo voláme pumpa na peniaze, a skúsime, ako sa vám to bude páčiť. Ak si myslíte, že stávka A je hodna 550 bodov [body boli po hre premenené na peniaze, ale nie v pomere 1:1], potom by ste mi za túto stávku boli ochotný dať 550 bodov...“
>
> …
>
> *Lichtenstein:* „Takže máte stávku A a ja poviem: ‚Ach, nechceli by ste radšej stávku B?‘“
>
> …
>
> *Pokusná osoba:* „Strácam peniaze.“
>
> *Lichtenstein:* „Ja od vás stávku B kúpim. Budem štedrý, dám vám viac ako 400 bodov. Dám vám 401 bodov. Ste ochotný predať mi stávku B za 401 bodov?“
>
> *Pokusná osoba:* „Samozrejme.“

258 Sarah Lichtenstein and Paul Slovic, „Reversals of Preference Between Bids and Choices in Gambing Decisions,“ *Journal of Experimental Psychology* 89, no. 1 (1971): 46–55.

259 William Poundstone, *Priceless: The Myth of Fair Value (and How to Take Advantage of It)* (Hill & Wang, 2010).

260 Sarah Lichtenstein and Paul Slovic, eds., *The Construction of Preference* (Cambridge University Press, 2006).

…

Lichtenstein: „Tak som zarobila 149 bodov.“

Pokusná osoba: „Takto dobre ja rozmýšľam. (smeje sa) Koľkokrát to budeme opakovať?“

…

Lichtenstein: „Myslím si, že som vás zatlačila tak ďaleko, ako sa dá bez toho, aby som vás urazila.“

Pokusná osoba: „To je pravda.“

Chce sa vám zakričať: „Už to *konečne vzdaj*! Intuícia *nie je vždy správna*!“

A potom je tu tá vec so zvláštnou hodnotou, ktorú ľudia pripisujú istote. Opäť, nemám tu svoje knihy, ale verím, že jeden experiment ukázal, že zmena z pravdepodobnosti 100 % na 99 % zavážila v mysliach ľudí viac než zmena z pravdepodobnosti 80 % na 20 %.

Problém s pripisovaním extra veľkej hodnoty istote je, že čo je *v jednom čase istota*, je *v inom čase pravdepodobnosť*.

V predchádzajúcej kapitole som hovoril o Allaisovom paradoxe:

- **1A.** Výhra 24 000 dolárov, naisto.
- **1B.** Šanca 33/34 vyhrať 27 000 dolárov, a šanca 1/34 nevyhrať nič.
- **2A.** Šanca 34% vyhrať 24 000 dolárov, a šanca 66% nevyhrať nič.
- **2B.** Šanca 33% vyhrať 27 000 dolárov, a šanca 67% nevyhrať nič.

Vzor naivných preferencií v Allaisovom paradoxe je 1A > 1B a 2B > 2A. Potom mi zaplatíte za prepnutie spínača z A na B, pretože by ste radšej mali šancu 33 % vyhrať 27 000 dolárov než šancu 34 % vyhrať 24 000 dolárov. Potom hod kocky odstráni časť masy pravdepodobnosti. V oboch prípadoch máte šancu *aspoň* 66 % nevyhrať nič. Tento hod kocky odstráni týchto 66 %. Takže teraz je možnosť B šanca 33/34 vyhrať 27 000 dolárov, ale možnosť A je *istota* vyhrania 24 000 dolárov. Ach, úžasná istota! Takže mi zaplatíte za prepnutie prepínača naspäť z B na A.

Keby som vám bol povedal už na začiatku, že toto všetko urobím, naozaj by ste mi zaplatili, aby som prepol ten prepínač a potom by ste mi zaplatili, aby som ho prepol naspäť? Alebo by ste si to radšej rozmysleli?

Kedykoľvek sa snažíte ohodnotiť zmenu pravdepodobnosti z 24 % na 23 % ako menej dôležitú než zmenu z ~1 na 99 0 – vždy keď sa snažíte ohodnotiť zmenu pravdepodobnosti vyššie, keď je pri konci škály – otvárate sa tomuto druhu vykorisťovania. Vždy viem nastaviť reťaz udalostí, ktoré odstránia masu pravdepodobnosti, po kúskoch, kým vám nezostane „istota“, ktorá otočí vaše preferencie. Čo je v jednom čase istota, je v inom čase neistota, a ak trváte na tom, že vzdialenosť medzi ~1 a 0,99 je špeciálna, môžem vaše preferencie v čase otočiť a vypumpovať z vás nejaké peniaze.

Môžem vás snáď presvedčiť, že toto je nerozumný vzorec?

Iste, ak tento blog nejaký čas čítate, uvedomujete si, že *vy* – samotný systém a proces, ktorý číta samotné tieto slová – ste chybný kus stroja. Vaše intuície vám nedávajú priamu, pravdivú informáciu o dobrom výbere. Ak tomu neveríte, je tu pár hazardných hier, ktoré by som si s vami rád zahral.

Existujú aj rôzne ďalšie hry, ktoré môžete s efektom istoty. Napríklad, ak niekomu ponúknete naisto 400 dolárov, alebo s pravdepodobnosťou 80 % 500 dolárov a s pravdepodobnosťou 20 % 300 dolárov, zvyčajne si vyberie tých 400 dolárov. Ale ak mu poviete, aby si predstavil, že má o 500 dolárov viac, a opýtate sa, či by radšej naisto stratil 100 dolárov, alebo s pravdepodobnosťou 20 % stratil 200 dolárov, zvyčajne si vyberie tú šancu stratiť 200 dolárov.[261] Rovnaká pravdepodobnostná distribúcia pre výsledky, rôzne opisy, rôzne voľby.

Áno, Virginia, naozaj by si *mala* skúšať násobiť užitočnosť výsledkov ich pravdepodobnosťou. Naozaj by si mala. Nehanbi sa používať čistú matematiku.

261 Kahneman and Tversky, „Prospect Theory: An Analysis of Decision Under Risk.“

V Allaisovom paradoxe si zisti, či 1 jednotka rozdielu medzi dostaním 24 000 dolárov a nedostaním ničoho preváži 33 jednotiek rozdielu medzi dostaním 24 000 dolárov a 27 000 dolárov. Ak áno, potom dávaj prednosť 1A pred 1B, a 2A pred 2B. Ak 33 jednotiek preváži 1 jednotku, uprednostňuj 1B pred 1A, a 2B pred 2A. Čo sa týka počítania užitočnosti peňazí, odporučil by som používať približný odhad predpokladajúci, že úžitok z peňazí je logaritmický. Ak už máte dosť peňazí, vyberte si B. Keby 24 000 zdvojnásobilo váš existujúci majetok, vyberte si A. Prípad 2 alebo prípad 1, to je jedno. Aha, a dajte si pozor, aby ste odhadovali užitočnosť celkovej hodnoty majetku – užitočnosť výsledných stavov sveta – nie *zmeny v* majetku, inak budete opäť nekonzistentní.

Mnoho včerajších komentátorov tvrdilo, že vzor preferencií nie je nerozumný, kvôli „užitočnosti istoty“ alebo niečomu podobnému. Jeden z komentátorov dokonca do rovnice očakávaného úžitku doplnil U(Istota).

Pamätá si niekto celú tú vec, že *očakávaný úžitok* a *úžitok* sú zásadne odlišné typy? Úžitok je z *výsledku*. Je to hodnota, ktorú pripisujete *konkrétnemu, pevne danému stavu sveta*. Nemôžete do funkcie úžitku doplniť pravdepodobnosť 1. To nedáva zmysel.

A skôr než si povzdychnete: „Hmmm... ty skrátka chceš, aby matematika bola pekná a uprat???“ pamätajte si, že v tomto prípade je cenou za odklonenie sa od bayesovskej cesty platenie niekomu za to, aby prepol prepínač a potom ho prepol naspäť.

Ale čo ten pevný hrejivý pocit bezpečia? Nie je aj *to* úžitok?

To je byť človekom. Ľudia nie sú maximalizátori očakávaného úžitku. Či si chcete oddýchnuť a zabávať sa, alebo platiť extra peniaze za pocit istoty, záleží na tom, či vám viac záleží na uspokojovaní svojich intuícií, alebo *na skutočnom dosahovaní cieľa*.

Ak chodíte do Las Vegas kvôli zábave, potom samozrejme nemyslite na očakávaný úžitok – aj tak idete prehrať svoje peniaze.

Ale čo keby namiesto 24 000 dolárov išlo o 24 000 ľudských životov? Účinok istoty je pri ľudských životoch ešte silnejší. Boli by ste ochotní zaplatiť jeden ľudský život za prepnutie prepínača, a potom ďalší za jeho prepnutie naspäť?

Tolerovaním obrátenia preferencií sa vysmievate z tvrdení, že optimalizujete. Ak cestujete zo San Jose do San Francisca a do Oaklandu a do San Jose, znovu a znovu, potom z toho môžete mať veľa hrejivých pocitov, ale nemôžete to interpretovať, že máte *cieľ* – že sa snažíte *niekam* dostať.

Keď máte kruhové preferencie, *nekormidlujete budúcnosť* – iba beháte v kruhu. Ak vás teší behanie samotné, nech sa páči. Ale ak máte cieľ – niečo, čo naozaj chcete dosiahnuť – obrátenie preferencií odhaľuje veľký problém. Prinajmenšom jedna z volieb, ktoré robíte, v skutočnosti neoptimalizuje budúcnosť v žiadnom koherentnom zmysle.

Ak vám záleží na hrejivom pocite istoty, nech sa páči. Ak ide o niekoho život, potom by ste si mali uvedomiť, že vaše intuície sú ako zamastený objektív, ktorým pozeráte na svet. Vaše pocity vám nedávajú priame, pravdivé informácie o strategických dôsledkoch – *pripadá* vám to tak, ale *nie je* to tak. Hrejivé pocity vás zavádzajú.

Existujú matematické zákony, ktoré riadia účinné stratégie na kormidlovanie budúcnosti. Keď ide o niečo *naozaj* dôležité – niečo dôležitejšie než vaše pocity šťastia z rozhodnutia – potom by ste sa mali starať o matematiku, ak vám na tom naozaj záleží.

* →

285. Morálne cítenie

Predstavte si, že nejaká choroba, alebo príšera, alebo vojna, alebo niečo zabíja ľudí. A predstavte si, že máte dostatok prostriedkov na to, aby ste uskutočnili iba jednu z nasledujúcich dvoch možností:

1. Zachrániť 400 životov, naisto.

→ http://lesswrong.com/lw/mz/zut_allais/

2. Zachrániť 500 životov s pravdepodobnosťou 90 %; nezachrániť žiaden život s pravdepodobnosťou 10 %.

Väčšina ľudí si vyberie možnosť 1. Čo je podľa mňa hlúpe; lebo ak vynásobíte 500 životov pravdepodobnosťou 90 %, dostanete očakávanú hodnotu 450 životov, čo je viac ako 400 v možnosti 1. (Marginálna užitočnosť zachránených životov neklesá, takže toto je primeraný výpočet.)

„Čože!“ zvoláte rozhorčene. „Ako môžeš hazardovať s ľudskými životmi? Ako môžeš myslieť na čísla, keď je v stávke tak veľa? Čo ak nastane pravdepodobnosť 10 % a každý zomrie? Dočerta s tvojou prekliatou logikou! Ty by si aj z okna skočil, keby to bolo rozumné!“

Ach, ale toto je zaujímavá vec. Ak tieto možnosti predložíte takto:

1. 100 ľudí zomrie, naisto.
2. Šanca 90 %, že nezomrie nikto; šanca 10 %, že zomrie 500 ľudí.

Potom si väčšina vyberie možnosť 2. *Napriek tomu, že je to ten istý hazard.* Vidíte, rovnako ako nám *istota* záchrany 400 životov *pripadá* omnoho príjemnejšia než neistý zisk, rovnako aj istú stratu *pociťujeme* horšie než neistú.

Môžete zaujať povýšené stanovisko aj pri tomto druhom opise: „Ako sa opovažuješ odsúdiť 100 ľudí na istú smrť, keď existuje taká dobrá šanca, že ich zachránime? Riziko budeme znášať spoločne! Dokonca aj keby bola iba šanca 75 %, že zachránime každého, stále by to stálo za to – dokiaľ je tam šanca – zachráni sa každý, alebo nikto!“

Viete čo? Tu nejde o vaše pocity. Ľudský život, so všetkými jeho radosťami a strasťami, nahromadenými počas desaťročí, je hoden viac než pocit pohody alebo nepohody vášho mozgu z nejakého plánu. Pripadá vám počítanie očakávaného úžitku príliš chladnokrvné na váš vkus? No, tak tento váš pocit nezaváži ani toľko čo pierko, keď sú v stávke životy. Držte hubu a počítajte.

Googol je 10^{100} – jednotka, za ktorou nasleduje sto núl. Googolplex je ešte nepochopiteľnejšie veľké číslo – je to 10^{googol}, jednotka, *za ktorou nasleduje googol núl*. Vezmite si nejaké triviálne nepohodlie, napríklad čkanie, a nejaké jednoznačne netriviálne nešťastie, napríklad byť postupne roztrhaný, končatinu za končatinou, sadistickými mutantnými žralokmi. Keby sme si museli vybrať, či zabrániť, aby sa googolplex ľuďom čkalo, alebo zabrániť, aby jedného človeka napadli žraloky, čo by ste si vybrali? Ak priradíte čkaniu *ľubovoľnú* zápornú hodnotu, potom pod hrozbou nekoherencie teórie rozhodovania musí existovať nejaké množstvo čkania, ktoré by sa dokopy vyrovnalo zápornej hodnote útoku žralokov. *Pre ľubovoľné konečné zlo* musí existovať nejaké množstvo čkania, ktoré by bolo ešte horšie.

Morálne dilemy, ako je táto, nie sú pojmové krvavé športy, existujúce iba nato, aby zabávali analytických filozofov na večierkoch. Sú to zjednodušené verzie tých druhov situácií, v ktorých sa v skutočnosti nachádzame každý deň. Mal by som minúť 50 dolárov na počítačovú hru, alebo darovať všetko na dobročinnosť? Mal by som zorganizovať zbierku na 700 000 dolárov na zaplatenie jednej transplantácie kostnej drene, alebo by som mal použiť tie isté peniaze na siete proti moskytom a zachrániť tak približne 280 detí pred smrťou na maláriu?

Napriek tomu sú mnohí, ktorí odmietajú vytrvalo rozmýšľať nad hojnosťou nepríjemných morálnych rozhodnutí v skutočnom svete – dokonca mnohí, ktorí sú hrdí, že toto odmietajú. Výskum ukazuje, že ľudia rozlišujú medzi „posvätnými hodnotami“, ako sú ľudské životy, a „obyčajnými hodnotami“, ako sú peniaze. Keď sa pokúsite vymeniť posvätnú hodnotu za obyčajnú hodnotu, pokusné osoby vyjadrujú veľké rozhorčenie. (Niekedy chcú potrestať osobu, ktorá to navrhla.)

Moja obľúbená anekdota v tomto smere pochádza od skupiny výskumníkov, ktorí vyhodnocovali efektivitu istého projektu, počítali náklady na jeden zachránený život, a odporučili vláde, že by sa tento projekt mal uskutočniť, pretože bol cenovo efektívny. Vládna agentúra odmietla túto správu, pretože, ako povedali, hodnotu ľudského života nemožno vyčísliť v dolároch. Po odmietnutí tejto správy sa agentúra rozhodla, že toto opatrenie *neuskutoční*.

Vymieňať posvätnej hodnoty za obyčajnú hodnotu je *naozaj hrozný pocit*. Iba vynásobiť úžitky by bolo príliš chladnokrvné – bolo by to, ako keby ste skočili z okna, keby to bolo rozumné...

Lenže altruizmus nie je to isté ako hrejivé pocity, ktoré máte z toho, že ste altruista. Ak to robíte pre duchovný úžitok, to nie je nič iné ako sebectvo. Prvoradý cieľ je pomôcť druhým, akokoľvek. Takže držte hubu a počítajte!

A ak sa vám zdá, že v tejto maximalizácii je niečo zúrivé, ako holý meč zákona, alebo žiara slnka – ak sa vám zdá, že v strede tejto rozumnosti je malý chladný plameň...

Nuž, ten druhý spôsob by mohol dosiahnuť, že by ste sa cítili lepšie. Ale nefungoval by.

A poviem vám toto: Ak odložíte svoju ľútosť za všetkým tým duchovným uspokojením, ktoré ste mohli mať – ak budete *celým srdcom* nasledovať Cestu, bez pocitu, že ste okrádaní – ak sa dokážete oddať rozumnosti bezo zvyšku, potom zistíte, že rozumnosť vám dáva naspäť.

Ale *táto* časť funguje iba vtedy, ak nechodíte okolo so slovami: „Cítil by som sa vo vnútri lepšie, keby som mohol byť menej rozumný.“ Mali by ste byť smutní, že máte príležitosť naozaj pomôcť ľuďom? Nemôžete dosiahnuť svoj plný potenciál, ak budete vnímať svoj dar ako bremeno.

286. „Intuície“ za „utilitariánstvom“

Kedysi som bol ohľadom metaetiky veľmi zmätený. Keď sa môj zmätok konečne vyjasnil, urobil som si zhrnutie svojich dovtedajších myšlienok. Zistil som, že moje morálne uvažovanie na konkrétnej úrovni bolo hodnotné, a moje morálne uvažovanie na metaúrovni bolo horšie než zbytočné. A toto vyzerá ako všeobecný syndróm – ľuďom sa omnoho lepšie darí pri diskusii, či je mučenie dobré alebo zlé, než pri diskusii o význame „dobra“ a „zla“. Preto mi pripadá prezieravé držať morálne diskusie na konkrétnej úrovni, kde sa len dá.

Občas ľudia namietajú voči diskusiám o morálke na základe toho, že morálka neexistuje, a namiesto preskakovania na zatiaľ neprebrané vysvetlenie toho, že slovo „existuje“ tu nie je tým správnym pojmom, vo všeobecnosti hovorím: „Ale aj tak, čo robíte?“ a vraciam tým diskusiu späť na konkrétnu úroveň.

Avšak Paul Gowder podotkol, že aj predstava, že máme uprednostniť $10^{10^{100}}$ zrniek prachu v $10^{10^{100}}$ očiach pred 50 rokmi mučenia jednej osoby, aj predstava „utilitariánstva“, závisia od „intuície“. Hovorí, že som tvrdil, že tieto dve nie sú zlučiteľné, ale obviňuje ma, že som neargumentoval za tie utilitariánske intuície, na ktoré sa odvolávam.

Nuž, „intuícia“ nie je to, ako by som opísal výpočty, na ktorých sa zakladá ľudská morálka a ktoré nás odlišujú, ako moralizátorov, od ideálneho filozofa dokonalej prázdnoty a/alebo kameňa. Ale som ochotný pracovne použiť slovo „intuícia“, pokiaľ si budeme pamätať, že „intuícia“ v tomto zmysle nie je v protiklade k rozumu, ale je to skôr kognitívny stavebný kameň, na ktorom sú postavené aj dlhé slovné argumenty, aj rýchle vnemové úsudky.

Projekt morálky vnímam ako projekt renormalizácie intuície. Máme intuície o veciach, ktoré vyzerajú želané alebo neželané, intuície o činoch, ktoré sú správne alebo nesprávne, intuície o tom, ako riešiť intuície, ktoré si navzájom odporujú, intuície o tom, ako konkrétne intuície systematizovať do všeobecných princípov.

Vymažte všetky tieto intuície, a nedostane vám ideálny filozof dokonalej prázdnoty, ale zostane vám kameň.

Ponechajte si všetky konkrétne intuície, ale odmietnite stavať na tých reflexívych, a nezostane vám ideálny dokonale spontánny a autentický filozof, ale chrochtajúci pračlovek, ktorý sa točí v kruhu kvôli cyklickým preferenciám a podobným nekonzistentnostiam.

„Intuícia“, ako ju tu pracovne používame, nie je nadávka, keď ide o morálku – z ničoho iného sa argumentovať nedá. Ešte aj modus ponens je v tomto zmysle „intuícia“ - akurát že modus ponens vyzerá ako dobrý nápad ešte aj potom, čo ho formalizujeme, zamyslíme sa nad ním, extrapolujeme ho, aby sme videli, či má rozumné dôsledky, a tak ďalej.

Takže to je „intuícia“.

Gowder však nepovedal, čo myslí slovom „utilitariánstvo“. Znamená utilitariánstvo, že...

- Správne konanie je jednoznačne dané dobrými dôsledkami?

- Chvályhodné konanie závisí od zdôvodniteľného očakávania dobrých dôsledkov?
- Pravdepodobnosti dôsledkov by sa mali normatívne diskontovať ich pravdepodobnosťou, čiže pravdepodobnosť 50 % niečoho zlého by mala mať v takýchto výpočtoch presne polovičnú váhu?
- Cnostné činy vždy zodpovedajú maximalizovaniu očakávaného úžitku, pri nejakej funkcii úžitku?
- Dve škodlivé udalosti sú horšie než jedna?
- Dva nezávislé prípady škody (nie tej istej osobe, navzájom sa neovplyvňujúce) sú presne dvakrát také zlé ako jeden?
- Pre ľubovoľné dve škody A a B, kde A je omnoho horšie ako B, existuje nejaká malá pravdepodobnosť taká, že riziko škody A s touto pravdepodobnosťou je prijateľnejšie než istota B?

Ak poviete, že niečo odporúčam, alebo že môj argument na niečo závisí a že to je nesprávne, prosím upresnite, čo táto vec je... mimochodom, súhlasím s 3, 4, 6 a 7, ale nie 4; nie som si istý formuláciou 1; a 2 je asi pravdivé, ale vyjadrené veľmi solipsistickým a sebeckým spôsobom: nemali by ste sa starať o to, či ste chvályhodní.

Aké sú teda tie „intuície“, od ktorých moje „utilitariánstvo“ závisí?

To je dosť hlboká téma, ale pokúsim sa do nej rýchlo pichnúť.

Po prvé, nie je to tak, že by mi niekto priniesol horeuvedený zoznam výrokov a ja by som sa rozhodol, ktoré z nich mi znejú „intuitívne“. Okrem iného, ak sa pokúsite porušiť „utilitariánstvo“, dostanete paradox, spory, kruhové preferencie, a ďalšie veci, ktoré nie sú ani tak príznakmi morálneho zla ako morálnej nekoherencie.

Keď budete chvíľu uvažovať o morálnych problémoch, a nájdete nové pravdy o svete, a dokonca objavíte znepokojujúce fakty o tom, ako fungujete vy sami, často skončíte s inými morálnymi názormi než ste začínali. Toto nie je celkom *definícia* morálneho pokroku, ale takto morálny pokrok prežívame.

Ako súčasť môjho prežívania morálneho pokroku som si pojmovo rozdelil otázky typu *Kam by sme mali ísť?* a otázky typu *Ako by sme sa tam mali dostať?* (Možno toto myslel Gowder slovami, že som „utilitarián“?)

Otázku, kam daná cesta vedie, dokážete odpovedať tak, že tú cestu prejdete a zistíte. Ak máte nepravdivú predstavu o tom, kam táto cesta vedie, táto nepravdivosť môže byť zničená pravdou veľmi priamym spôsobom.

Čo sa týka toho, či chcete ísť na konkrétne miesto, táto túžba nie je celkom imúnna proti ničivej sile pravdy. Mohli by ste tam ísť a dodatočne zistiť, že to ľutujete (čo nie je definícia morálnej chyby, ale takto morálnu chybu prežívame).

Ale aj tak sa oplatí rozlišovať medzi tým, či konkrétne miesto vyzerá dobre a či chceme ísť konkrétnou cestou na konkrétne miesto.

Naše intuície ohľadom toho, kam chceme ísť, sú dosť pochybné, ale naše intuície ohľadom toho, ako sa tam dostaneme, sú úprimne povedané, na figu. Keď vám dvestoosemdesiatasiedma výskumná štúdia ukáže, že ľudia si odseknú vlastnú nohu, pokiaľ problém opíšete z nesprávneho uhla, začnete pochybovať o svojich dojmoch.

Keď ste čítali dosť o necitlivosti voči rozsahu – ľudia zaplatia iba o 28 % viac za ochranu všetkých prírodných rezervácií v Ontariu než za jednu rezerváciu, ľudia zaplatia rovnaké množstvo za záchranu 50 000 životov ako za 5 000 životov... a takéto veci...

No, najhorší prípad necitlivosti voči rozsahu, o akom som kedy počul, opísal Slovic tu:

> Iné nedávne výskumy ukazujú podobné výsledky. Dvaja izraelskí psychológovia žiadali ľudí, aby prispeli na drahú liečbu potrebnú na záchranu života. Mohli tento príspevok dať skupine ôsmych chorých detí, alebo jedinému dieťaťu vybranému z tejto skupiny. Cieľová suma potrebná na záchranu dieťaťa (alebo všetkých detí) bola v oboch prípadoch rovnaká.

> Príspevky na jednotlivého člena skupiny výrazne prevyšovali príspevky na skupinu ako celok.[262]

Existujú podobné výskumy s rovnakou pointou, ale ja vám tu ukazujem iba jeden príklad, pretože, iste pochopíte, osem príkladov by na vás pravdepodobne urobilo menší dojem.

Ak poznáte základnú paradigmu týchto pokusov, potom je dôvod pre opísané správanie celkom zrejmý – sústrediť svoju pozornosť na jediné dieťa vyvoláva silnejšie emocionálne vzrušenie ako snažiť sa pozornosť rozdeliť medzi osem detí zároveň. Ľudia sú preto ochotní zaplatiť viac za pomoc jednému dieťaťu než za pomoc ôsmim.

Teraz by ste sa mohli pozrieť na túto intuíciu a pomyslieť si, že odhaľuje nejakú neuveriteľne hlbokú morálnu pravdu, ktorá ukazuje, že šťastie jedného dieťaťa je akosi znehodnotené šťastím ostatných detí.

Ale čo miliardy ďalších detí na tomto svete? Prečo nepovažujeme za zlý nápad pomôcť tomuto jednému dieťaťu, ak to spôsobuje, že hodnota všetkých zvyšných detí klesne? Ako môže byť výrazne lepšie mať 1 329 342 410 šťastných detí než 1 329 342 409, ale z nejakého dôvodu je horšie mať ich o sedem viac, čiže 1 329 342 417?

Alebo sa na to môžete pozrieť a povedať: „Táto intuícia je nesprávna: mozog nedokáže úspešne vynásobiť ôsmimi a dostať väčšie množstvo než mal na začiatku. Ale mal by, normatívne povedané.“

A keď si raz uvedomíte, že mozog nedokáže násobiť ôsmimi, potom sa vám prestane zdať, že tie ďalšie prípady ignorovania rozsahu odhaľujú nejakú zásadnú pravdu o tom, že 50 000 životov sp zaslúži rovnaké úsilie ako 5 000 životov, alebo niečo také. Prestanete mať dojem, že sa pozeráte na zjavenia nejakých hlbokých morálnych právd o nespočítateľných úžitkoch. To len ten mozog nedokáže sakra násobiť. Čísla sú vyhodené z okna.

Ak môžete minúť 100 dolárov a miniete po 20 z nich na každé z 5 úsilí zachrániť 5 000 životov, urobíte horšie než keby ste minuli 100 dolárov na jedno úsilie zachrániť 50 000 životov. Podobne je to, ak takéto rozhodnutia urobí 10 rôznych ľudí, nie jedna osoba. Akonáhle začnete veriť, že je lepšie zachrániť 50 000 životov než 25 000 životov, táto jednoduchá preferencia konečného cieľa má dôsledok pre voľbu ciest, keď zvažujete päť rôznych udalostí, ktoré zachránia 5 000 životov.

(Je všeobecným princípom, že bayesiánci nevidia rozdiel medzi dlhodobo správnou odpoveďou a krátkodobo správnou odpoveďou; nikdy nedostanete dve rôzne odpovede pri počítaní tej istej otázky dvoma rôznymi spôsobmi. Ale dlhodobá správnosť je užitočná pumpa intuície, preto o nej aj tak budem hovoriť.)

Agregačná hodnotiaca stratégia „drž hubu a počítaj“ vzniká z jednoduchej preferencie mať niečoho viac – zachrániť toľko životov, koľko sa dá – keď máte opísať všeobecné princípy na vyberanie si viac než jedenkrát, konanie viac než jedenkrát, plánovanie viac než jedenkrát.

Agregácia vzniká aj z tvrdenia, že lokálny výber, či zachrániť jeden život, nezávisí od toho, koľko životov už existuje niekde ďaleko na opačnej strane planéty alebo ďaleko na opačnej strane vesmíru. Tri životy sú jeden a jeden a jeden. Nezáleží na tom, koľko miliárd je na tom lepšie, alebo je na tom horšie. 3 = 1 + 1 + 1, bez ohľad na to, aké ďalšie množstvá pripočítate k obom stranám rovnice. A keď pridáte ďalší život, dostanete 4 = 1 + 1 + 1 + 1. Toto je agregácia.

Keď ste čítali dosť výskumov o heuristík a skresleniach, a dosť dôkazov koherencie a jedinečnosti bayesovských pravdepodobností a očakávaného úžitku, a videli ste efekty „Dutch book“ a „pumpy na peniaze“, ktoré trestajú zaobchádzanie s neistými výsledkami iným spôsobom, potom nevnímate obrátenie preferencií v Allaisovom paradoxe ako odhalenie nejakej neuveriteľne hlbokej morálnej pravdy o vnútornej hodnote istoty. Iba vám to ukazuje, že mozog nevie sakra násobiť.

Primitívne vnemové intuície, vďaka ktorým máte z voľby „dobrý pocit“, nezvládajú pravdepodobnostné cesty časom veľmi šikovne, najmä keď boli pravdepodobnosti vyjadrené symbolicky

262 Paul Slovic, „Numbed by Numbers,“ *Foreign Policy* (March 2007), http://foreignpolicy.com/2007/03/13/numbed-by-numbers/.

a nie ako frekvencie. Takže reflektujete, vytvoríte si dôveryhodnejšiu logiku, a premyslíte si to pomocou slov.

Keď vidíte, ako ľudia trvajú na tom, že absolútne žiadne množstvo peňazí nemá takú hodnotu ako jeden ľudský život, a potom cestujú jednu míľu navyše aby ušetrili 10 dolárov; alebo keď vidíte, že ľudia trvajú na tom, že žiadne množstvo peňazí nestojí za zníženie zdravia, a potom si vyberú najlacnejšie možné zdravotné poistenie; potom si nemyslíte, že ich protesty vyjadrujú nejakú hlbokú pravdu o nesúmerateľných úžitkoch.

Zčasti ide samozrejme o to, že primitívne intuície úspešne neznižujú emocionálny dopad symbolov vyjadrujúcich malé sumy – všetko, o čom hovoríte, vyzerá ako „množstvo hodné úvahy“.

A zčasti to súvisí s uprednostňovaním bezpodmienečných spoločenských pravidiel pred podmienečnými spoločenskými pravidlami. Podmienečné pravidlá sa zdajú slabšie, ľahšie manipulovateľné. Ak existuje nejaká výnimka, ktorá vláde dovolí legálne mučiť ľudí, potom vláda cez túto dieru pretlačí celý kamión.

Takže sa zdá, že by mali existovať bezpodmienečné spoločenské zákazy proti uprednostňovaniu peňazí pred životom, za ktorými nenasleduje žiadne „ale“. Ani len „ale tisíc dolárov nie je hodno šance na záchranu života s pravdepodobnosťou 0,000 000 000 1 %.“ Hoci táto voľba sa samozrejme prejaví vždy keď si kýchneme bez zavolania lekára.

Rétorika posvätnosti dostáva bonusové body za to, že vyzerá ako vyjadrenie *neobmedzeného* záväzku, ako *bezpodmienečné* odmietnutie, ktoré signalizuje dôveryhodnosť a odmietnutie kompromisov. Dôjdete teda k záveru, že morálna rétorika zastáva kvalitatívne rozdiely, pretože zastávanie kvantitatívnych výmen by znelo ako plánovanie podrazu.

V takýchto prípadoch ľudia energicky chcú vyhodiť kvantity z okna, a rozčuľujú sa, ak sa pokúšate kvantity vrátiť, pretože kvantity znejú ako podmienky, ktoré by oslabili pravidlo.

Lenže vy nedôjdete k záveru, že *v skutočnosti* existujú dve úrovne úžitku s lexikálnym usporiadaním. Nedôjdete k záveru, že naozaj existuje nekonečne strmý morálny gradient, nejaký atóm, ktorý sa pohne o Planckovu vzdialenosť (v našom spojitom fyzikálnom vesmíre) a pošle úžitok z nuly do nekonečna. Nedôjdete k záveru, že úžitok treba vyjadrovať pomocou hyperreálnych čísel. Pretože tá nižšia vrstva by jednoducho v každej rovnici zmizla. Nikdy by nebola hodna toho najmenšieho úsilia spočítať ju. Všetky rozhodnutia by určovala tá vyššia vrstva, a každú myšlienku by sme trávili iba rozmýšľaním o tejto vyššej vrstve, keby tá vyššia vrstva naozaj mala lexikálnu prioritu.

Ako raz podotkol Peter Norvig, keby Asimovovi roboti mali *prísnu* prioritu Prvého zákona robotiky („Robot nesmie ublížiť človeku, ani dopustiť, aby vďaka jeho nečinnosti bolo človeku ublížené“), potom by správanie robotov nikdy nevykazovalo žiadnu známku ďalších dvoch zákonov; vždy by existoval nejaký maličký faktor prvého zákona, ktorý by stačil na určenie rozhodnutia.

Ak sa na nejakú hodnotu oplatí čo len myslieť, musí sa oplatiť vymieňať ju za všetky ostatné hodnoty, na ktoré sa oplatí myslieť, pretože samotná myšlienka je obmedzený zdroj, s ktorým treba hospodáriť. Keď odhaľujete hodnotu, odhaľujete úžitok.

Nehovorím, že by morálka mala vždy byť jednoduchá. Už som povedal, že zmyslom hudby je viac než samotné šťastie, viac než iba rozsvietenie centra radosti. Radšej by som videl, aby hudbu skladali ľudia než nevedomé algoritmy strojového učenia, aby niekto mal radosť zo skladania; zaujímam sa o samotnú cestu, nielen o jej cieľ. A som pripravený počuť, ak mi poviete, že hodnota hudby je hlbšia, a zahŕňa viac komplikácií než si uvedomujem – že hodnotenie tejto jednej udalosti je zložitejšie než viem.

Ale to sa týka *jednej udalosti*. Keď ide o násobenie kvantít a pravdepodobností, treba sa komplikáciám vyhýbať – prinajmenšom ak vám viac záleží na cieli než na samotnej ceste. Keď ste reflektovali dosť intuícií a opravili dosť absurdít, začnete vidieť spoločný menovateľ, fungujúci metaprincíp, ktorý možno vyjadriť ako „drž hubu a počítaj“.

Keď ide o hudbu, záleží mi na samotnej ceste.

Keď ide o životy, držím hubu a počítam.

Je dôležitejšie, aby boli zachránené životy, než aby sme sa pridŕžali nejakého konkrétneho rituálu zachraňovania. A optimálna cesta k tomuto cieľu sa riadi zákonmi, ktoré sú jednoduché, lebo sú matematické.

A to je dôvod, prečo som utilitarián – prinajmenšom keď robím niečo, čo je omnoho dôležitejšie než moje pocity ohľadom toho – čo je väčšinou, pretože utilitariánov je málo a neurobenej práce je veľa.

*→

287. Účel nesvätí prostriedky (u ľudí)

> Ak ciele neospravedlňujú prostriedky, tak čo teda?
>
> --uvádzané rôzne zdroje

> Predstavujem si, že fungujem na nepriateľskom hardvéri.
>
> --Justin Corwin

Ľudia si mohli vyvinúť štruktúru politickej revolúcie, kde na začiatku veria, že sú morálne nadradení skazenej súčasnej mocenskej štruktúre, ale nakoniec moc skazí ich samotných -- nie vďaka nejakému plánu v ich vlastnej mysli, ale ako ozvena predkov, ktorí urobili to isté a vďaka tomu sa rozmnožili.

Toto zapadá do šablóny:

> V niektorých prípadoch sa ľudia vyvinuli takým spôsobom, že si myslia, že robia X kvôli spoločensky prospešnému dôvodu Y, ale keď ľudia v skutočnosti robia X, vykonajú sa ďalšie adaptácie, ktoré prinesú osobne prospešný dôsledok Z.

Odtiaľto prejde k svojmu hlavnému bodu, k otázke, ktorá je *značne* mimo oblasti klasickej bayesovskej teórie rozhodovania:

> Čo ak fungujem na skazenom hardvéri?

V takomto prípade sa môžete pristihnúť, ako hovoríte také napohľad paradoxné výroky -- čisté nezmysly z pohľadu klasickej teórie rozhodovania -- ako:

> Cieľ neospravedlňuje prostriedky.

Lenže ak fungujete na skazenom hardvéri, potom ak sebareflexiou zistíte, že sa vám *zdá* ako spravodlivý a altruistický čin uchopiť moc do vlastných rúk... takéto *zdanie* nemusí byť veľkou indíciu v prospech výroku, že uchopenie moc do vlastných rúk je naozaj to, čo najviac prospeje vášmu kmeňu.

Silou naivného realizmu vám skazený hardvér, na ktorom fungujete, a skazené dojmy, ktoré vám počíta, budú pripadať ako tkanivo samotného skutočného sveta -- jednoducho, že takto sa veci majú.

A preto máme napohľad bizarné pravidlo: „Pre dobro tvojho kmeňa, nepodvádzaj aby si sa chopil moci *ani vtedy, keď by to prinieslo celkový úžitok tvojmu kmeňu*.“

Naozaj je múdrejšie formulovať to týmto spôsobom. Keby ste iba povedali: „keď sa *zdá*, že by to prinieslo celkový úžitok tvojmu kmeňu“, dostali by ste ľudí, ktorí hovoria: „Ale to nie je iba nejaké *zdanie* -- to by *naozaj* prinieslo celkovú úžitok nášmu kmeňu, keby som sa ja chopil moci.“

Predstava nedôveryhodného hardvéru vyzerá ako niečo úplne mimo oblasti klasickej teórie rozhodovania. (Čo to robí s reflexívnou teóriu rozhodovania, neviem zatiaľ povedať, ale zdá sa, že to by bola vhodná úroveň na jej riešenie.)

Lenže na ľudskej úrovni takáto záplata vyzerá samozrejme. Akonáhle viete o nejakom pokrivení, vytvoríte pravidlá popisujúce toto pokrivené správanie a zakážete ho. Pravidlo, ktoré hovorí: „Pre dobro nášho kmeňa, nepodvádzaj aby si sa chopil moci, dokonca ani pre dobro nášho kmeňa.“ Alebo: „Pre dobro nášho kmeňa, nevraždi, dokonca ani pre dobro nášho kmeňa.“

→ http://lesswrong.com/lw/n9/the_intuitions_behind_utilitarianism/

A teraz príde filozof a predloží svoj „myšlienkový experiment“ -- nastaví scenár, v ktorom je *stanovené*, že *jediný* možný spôsob, ako zachrániť päť nevinných životov, je zavraždiť jedného nevinného človeka, a že táto vražda *naisto* zachráni tých päť životov. „Vlak sa rúti smerom na päť nevinných ľudí, ktorých nie je možné varovať, aby mu uskočili z dráhy, ale môžeš strčiť jedného nevinného človeka do dráhy tohto vlaku, čím sa tento vlak zastaví. Toto je jediná možnosť, ktorú máš; čo urobíš?“

Altruistický človek, ktorý prijal určité deontologické zákazy -- ktoré vyzerajú dobre zdôvodnené niektorými historickými štatistikami o dôsledkoch rozmýšľania určitými spôsobmi na nedôveryhodnom hardvéri -- môže prežívať nejaké duševné rozrušenie, keď sa stretne s týmto myšlienkovým experimentom.

Tu je teda odpoveď na scenár daného filozofa, ktorú som zatiaľ ešte nepočul od žiadnej filozofovej obete:

> Stanovil si, že *jediný možný* spôsob, ako zachrániť päť nevinných životov, je zavraždiť jedného nevinného človeka, a že táto vražda *naisto* zachráni tých päť životov, a že tieto fakty sú mi naozaj stopercentne *známe*. Lenže keďže ja fungujem na skazenom hardvéri, nie je možné, aby som sa nachádzal v *epistemickom stave*, ktorý odo mňa žiadaš, aby som si predstavil. Preto odpovedám, že v spoločnosti Umelých Inteligencií, ktoré by mali osobnosť a nemali by žiaden zabudovaný sklon byť skazené mocou, bolo by správne, aby UI zavraždila jednu nevinnú osobu, aby tým zachránila päť, a všetci jej rovesníci by súhlasili. Ja však odmietam vztiahnuť túto odpoveď na seba, pretože epistemický stav, ktorý odo mňa žiadaš, aby som si predstavil, môže existovať iba medzi inými druhmi bytostí, nie medzi ľuďmi.

Dobre, toto mi pripadá ako vyhýbanie sa. Myslím si, že vesmír je dostatočne neláskavý na to, aby sme boli právom nútení zvažovať situácie tohto druhu. Ten typ človeka, ktorý chodí a navrhuje druhým podobné myšlienkové experimenty, si možno naozaj zaslúži takýto typ odpovede. Lenže každý ľudský právny systém stelesňuje nejakú odpoveď na otázku: „Koľko nevinných ľudí môžeme dať do väzenia, aby sme dostali tých vinných?“, dokonca aj keď toto číslo nie je nikde napísané.

Ako človek sa snažím dodržiavať deontologické zákazy, ktoré ľudia vytvorili, aby žili navzájom v mieri. Ale nemyslím si, že naše deontologické zákazy sú *doslova sami osebe bez ohľadu na dôsledky definitívne správne*. Schvaľujem princíp „účel nesvätí prostriedky“ ako princíp, ktorým sa majú riadiť ľudia fungujúci na skazenom hardvéri, ale neschvaľoval by som ho ako princíp pre spoločnosť UI, ktoré robia dobre kalibrované odhady. (Pokiaľ máte jednu UI v ľudskej spoločnosti, to prináša ďalšie otázky, napríklad či ľudia budú napodobňovať jej správanie.)

Preto by som nepovedal, že dobre navrhnutá Priateľská UI musí nevyhnutne odmietnuť zhodiť toho jedného človeka z útesu, aby tým zastavila vlak. Samozrejme by som očakával, že každá slušná superinteligencia vymyslí nejakú tretiu, ešte lepšiu možnosť. Ale keby toto boli jediné dve možnosti, a PUI usúdi, že je múdrejšie zhodiť toho jedného človeka z útesu -- dokonca aj po zohľadnení následných účinkov na ľudí, ktorí vidia, ako sa to stalo, a budú túto informáciu šíriť, atď. -- potom by som nepovažoval za dôvod na poplach, ak *UI* povie, že správna vec je obetovať jedného, aby sa zachránili piati. Opäť, *ja* osobne nechodím strkať ľudí do dráhy vlakom, ani nekradnem z bánk, aby som financoval svoje altruistické projekty. *Ja* som totiž človek. Ale pre Priateľskú UI by skazenosť mocou bola ako keby z nej začala tiecť červená krv. Sklon byť skazený mocou je konkrétna biologická adaptácia, podporovaná konkrétnymi kognitívnymi obvodmi, ktorú do nás zabudovali gény z jasného evolučného dôvodu. Neobjavila by sa spontánne v kóde Priateľskej UI rovnako ako by jej tranzistory nezačali krvácať.

Šiel by som ešte ďalej a povedal, že keby ste mali mysle so zabudovaným pokrivením, ktoré by spôsobovalo, že by *preceňovali* vonkajšie škodlivé dôsledky činov, ktoré im osožia, potom by potrebovali pravidlo „účel nezakazuje prostriedky“ -- že by ste mali robiť to, čo pomáha vám samotným, dokonca aj keď (sa zdá, že) to škodí vášmu kmeňu. Podľa tejto hypotézy, keby ich spoločnosť nemala toto pravidlo, tieto mysle by odmietli dýchať vzduch zo strachu, že spotrebujú kyslík niekoho iného, a všetky by zomreli. Im by sa občasné prestrelenie, pri ktorom sa jedna osoba zmocní osobného prospechu na úkor spoločnosti, zdalo ako rovnaká cnosť opatrnosti -- a vskutku by to *bola* rovnaká cnosť opatrnosti -- ako

keď sa jeden z nás, ľudí, opatrne vzdá príležitosti ukradnúť peceň chleba, ktorý by mu naozaj priniesol viac úžitku než by priniesol škody predavačovi (vrátane následných účinkov).

„Cieľ neospravedlňuje prostriedky“ je iba konsekvencialistické uvažovanie o jednu meta-úroveň vyššie. Keby si ľudia začali na *objektovej* úrovni myslieť, že cieľ ospravedlňuje prostriedky, malo by to vzhľadom na naše nedôveryhodné mozgy hrozné dôsledky; preto by ľudia takto myslieť nemali. Ale toto celé je stále v konečnom dôsledku konsekvencializmus. Akurát je to *reflexívny* konsekvencializmus, pre bytosti, ktoré vedia, že ich okamžité rozhodovanie sa deje na nedôveryhodnom hardvéri.

*→

288. Etické zákazy

> Zabíjali by ste malé deti, keby to bola správna vec? Ak nie, za akých okolností by ste neboli ochotní urobiť správnu vec? Ak áno, nakoľko správna vec by to musela byť vzhľadom na aký počet detí?
>
> --hrozná otázka z pracovného pohovoru

Keď teraz na chvíľu zmením rolu, z *profesionálneho* hľadiska ma zaujíma teória rozhodovania ohľadom „vecí, ktoré by ste nemali robiť ani keby to bola tá správna vec“.

Predstavme si, že máme reflexívnu UI, sebamodifikujúcu a sebazdokonaľujúcu, v nejakom medzistupni procesu vývoja. Konkrétne, systém cieľov tejto UI nie je hotový -- tvar jeho motivácií sa ešte stále načítava, učí, testuje, alebo dolaďuje.

Hej, videl som už mnoho spôsobov, ako dokašľať návrh systému cieľov UI, ktorých výsledkom je systém rozhodovania, ktorý sa vzhľadom na svoje ciele rozhodne, že vesmír treba vyplniť malými molekulárnymi smajlíkmi alebo niečím podobným. Vo všeobecnosti majú tieto smrtiace odporučenia zároveň tú vlastnosť, že UI nebude túžiť po tom, aby ju jej programátori opravili. Ak je daná UI *dostatočne* rozvinutá -- čo môže byť už aj v nejakom medzistupni vývoja -- potom si táto UI môže uvedomiť, že ak podvedie svojich programátorov, ak ukryje zmeny vo svojich myšlienkach, pomôže jej to premeniť vesmír na smajlíkov.

Z pohľadu nás ako programátorov, *ak nastala skutočnosť*, že sa UI rozhodla ukryť svoje myšlienky pred programátormi, alebo inak vedome konať, aby nás oklamala, potom vyzerá pravdepodobne, že sa v jej systéme cieľov vyskytol nejaký nezamýšľaný dôsledok. Považovali by sme za pravdepodobné, že daná UI *nefunguje* tak, ako bolo zamýšľané, ale že sme pravdepodobne nejako domotali funkciu úžitku danej UI. Takže teraz chce UI premeniť vesmír ma malé počítadlá systému odmien, alebo niečo podobné, a teraz má motív to pred nami ukrývať.

Nuž, predpokladajme, že *nejdeme* implementovať nejakú Skvelú Myšlienku na objektovej úrovni ako funkciu úžitku danej UI. Namiesto toho urobíme niečo pokročilejšie a rekurzívne -- vytvoríme systém cieľov, ktorý rozoznáva programátorov tam vonku (a záleží mu na nich). Systém cieľov, ktorý pomocou nejakej netriviálnej vnútornej štruktúry „vie, že je programovaný“ a „vie, že je neúplný“. Potom by ste možno mohli mať a udržať si pravidlo:

> Ak (sa rozhodnem, že) oklamať mojich programátorov je tá správna vec, vykonaj kontrolované vypnutie (namiesto urobenia tej správnej veci).

A daná UI by si udržala toto pravidlo, aj napriek prerábkam vlastného kódu sebamodifikujúcou UI, pretože vo svojom štruktúrne netriviálnom systéme cieľov súčasná UI rozumie, že toto rozhodnutie u budúcej UI *pravdepodobne* naznačuje niečo definované-ako-zlyhanie. Navyše, súčasná UI vie, že keby sa budúca UI pokúsila *vyhodnotiť* úžitok z vykonania vypnutia, akonáhle už toto hypotetické zlyhanie nastalo, budúca UI by sa pravdepodobne *rozhodla* nevypnúť sa. Preto by sa toto vypnutie malo uskutočniť bezpodmienečne, automaticky, bez toho, aby systém cieľov dostal druhú šancu zmeniť výpočet, čo je správna vec.

→ http://lesswrong.com/lw/uv/ends_dont_justify_means_among_humans/

Nepôjdem teraz do temných hlbín presnej matematickej štruktúry, pretože to by bolo mimo rozsahu tohto blogu. Ani zatiaľ nepoznám temné hlbiny tejto matematickej štruktúry. Vyzerá to, že by to *malo* byť možné, ak robíte veci, ktoré sú pokročilé a rekurzívne a majú netriviálnu (ale konzistentnú) štruktúru. Ale zatiaľ som túto úroveň ešte nedosiahol, takže zatiaľ je to iba sen.

Lenže naša téma teraz nie je pokročilá UI, ale ľudská etika. Scenár s UI som uviedol, aby som jasnejšie vysvetlil čudnú myšlienku *etických zákazov*:

> Nemali by ste nikdy zavraždiť nevinného človeka, ktorý vám pomohol, *dokonca ani keby to bola tá správna vec*; pretože je omnoho väčšia pravdepodobnosť, že *sa mýlite*, než že je zavraždenie nevinného človeka, ktorý vám pomohol, tá správna vec.

Znie to rozumne?

Počas 2. svetovej vojny sa stalo nevyhnutné zničiť nemecké zásoby deutéria, spomaľovača neutrónov, aby sa zastavili nemecké pokusy dosiahnuť štiepnu jadrovú reakciu. Ich dodávky deutéria v tomto bode prichádzali z dobytého zariadenia v Nórsku. Zásielka ťažkej vody bola na palube nórskej trajektovej lode SF Hydro. Knut Haukelid a traja ďalší sa prešmykli na palubu lode, aby ju sabotovali, ale tam sabotérov odhalil strážca lode. Haukelid mu povedal, že utekajú pred gestapom, a strážca okamžite súhlasil, že si nebude všímať ich prítomnosť. Haukelid „zvažoval, či by nemal varovať ich dobrodinca, ale rozhodol sa, že by to mohlo ohroziť ich poslanie, a tak mu iba poďakoval a potriasol ruku.“[263] A tak sa civilná trajektová loď *Hydro* potopila v najhlbšej časti jazera; osemnásť členov posádky zomrelo, dvadsať deväť bolo zachránených. Niektorí nórski záchranári mali pocit, že by prítomných nemeckých vojakov mali nechať utopiť sa, ale tento postoj bol v menšine, a tak štyroch Nemcov zachránili. A toto bolo prakticky koniec nacistického programu atómových zbraní.

Dobrý ťah? Zlý ťah? Nemecko by *s veľkou pravdepodobnosťou* tak či tak bombu nebolo získalo... Absolútne zúfalo dúfam, že nikdy nebudem čeliť takejto voľbe, ale v konečnom dôsledku voči nej nemôžem povedať jediné slovo.

Na druhej strane, pokiaľ ide o pravidlo:

> Nikdy sa nepokúšaj podviesť sám seba, ani si nedávaj dôvod veriť niečomu inému než je pravdepodobne pravda; pretože aj keď prídeš s úžasne chytrým dôvodom, je omnoho pravdepodobnejšie, že si urobil chybu, než že môžeš rozumne očakávať, že toto bude z dlhodobého hľadiska čistý zisk.

Potom naozaj *nepoznám* nikoho, kto bol vedome postavený pred výnimku. Existujú situácie, keď sa pokúšate presvedčiť sami seba „Neskrývam vo svojej pivnici žiadnych židov“ predtým ako sa porozprávate s dôstojníkom Gestapa. Ale aj vtedy stále poznáte pravdu, iba sa pokúšate vytvoriť niečo ako alternatívne ja, ktoré existuje iba vo vašej predstavivosti, fasádu na rozprávanie sa s dôstojníkom Gestapa.

Ale aby ste naozaj verili niečomu, čo nie je pravda? Neviem, či niekedy existoval niekto, pre koho to bol *poznateľne* dobrý nápad. Som si istý, že v ľudskej histórii bolo veľa prípadov, keď pre človeka X bolo lepšie mať nepravdivý názor Y. A podobne, v každom losovaní je vždy nejaká množina vyhrávajúcich čísel žrebov. Akurát *vedieť, ktorý žreb v lotérii vyhrá*, je tá epistemicky náročná časť, podobne ako je pre X vedieť, kedy je preňho lepšie mať nepravdivý názor.

Sebaklamy sú tie najhoršie druhy stávok na čierne labute, omnoho horšie než lži, pretože ak nepoznáte skutočný stav vecí, nedokážete ani odhadnúť, aký bude trest za váš sebaklam. Stačí, aby to vybuchlo iba raz a môžete prísť o všetko dobré, čo to kedy urobilo. Jediný raz, keď sa budete modliť k Bohu po objavení hrčky, namiesto aby ste išli k lekárovi. To stačí na to, aby ste prišli o život. Všetko to šťastie, ktoré hrejivá myšlienka posmrtného života kedy vytvorila v ľudstve, bolo teraz viac než vyvážené neschopnosťou ľudstva zaviesť systematickú kryonickú konzerváciu potom, čo sa priemyselná výroba tekutého dusíka stala lacnou. A nemyslím si, že mal niekto niekedy na mysli takýto druh zlyhania ako možnú katastrofu, keď hovoril: „Ale my potrebujeme náboženskú vieru, aby sme zmiernili strach zo smrti.“ O tomto všetkom sú stávky na čierne labute -- nečakané katastrofy.

263 Richard Rhodes, *The Making of the Atomic Bomb* (New York: Simon & Schuster, 1986).

Možno sa vám prepečia jedna alebo dve stávky na čierne labute -- nie *vždy* vás dostanú. Tak to urobíte znova, a potom katastrofa príde a preváži všetky výhody a ešte niečo navyše. O tomto všetkom sú stávky na čierne labute.

Preto je ťažké vedieť, kedy je bezpečné veriť lži (za predpokladu, že sa vôbec dokážete takto mentálne ohnúť) -- časť podstaty stávok na čierne labute je, že nevidíte tú guľku, ktorá vás zabije; a keďže naše vnímanie jednoducho vyzerá ako spôsob, ako sa veci majú, tak to vyzerá, že tam žiadna guľka nie je, bodka.

Preto by som povedal, že existuje etický zákaz proti sebaklamu. Nazývam to „etický zákaz“ ani nie preto, že je to otázka medziľudskej morálky (hoci je), ale preto, lebo je to pravidlo, ktoré vás chráni pred vašou vlastnou chytrosťou -- ovládanie pokušenia urobiť to, čo *vyzerá ako* správna vec.

Teraz teda máme dva druhy situácií, ktoré dokážu podporiť „etický zákaz“, pravidlo nerobiť niečo ani keď je to tá správna vec. (To jest, zdržíte sa „aj keď váš mozog vypočítal, že je to tá správna vec“, ale toto vám samozrejme bude skrátka *pripadať ako* „tá správna vec“.)

Po prvé, sme ľudia a fungujeme na skazenom hardvéri, preto môžeme zovšeobecňovať druhy situácií, v ktorých keď poviete napríklad „Nastal čas vykradnúť nejaké banky pre väčšie dobro,“ považujeme za omnoho pravdepodobnejšie, že ste skazení, než že je to naozaj tak. (Všimnite si, že netvrdíme, že to tak *nikdy* nemôže byť *naozaj*, iba spochybňujeme *epistemický* stav, v ktorom ste si *zdôvodnili, že môžete dôverovať* svojmu výpočtu, že je to tá správna vec -- žreby v poctivej lotérii môžu vyhrať, ale vy si nemôžete zdôvodniť ich nákup.)

Po druhé, história nás môže naučiť, že isté druhy činov sú stávky na čierne labute, čiže že niekedy katastrofálne vybuchnú z dôvodov, ktoré sa nenachádzajú v modeli rozhodujúceho sa. Takže aj keď si v rámci modelu vypočítame, že niečo vyzerá ako tá správna vec, použijeme ďalšie vedomosti o probléme čiernych labutí, aby sme dospeli k zákazu tejto veci.

Ale určite... ak si človek *tieto dôvody uvedomuje*... potom môže jednoducho prerobiť tieto výpočty, zohľadniť to v nich. Môžeme teda vykrádať banky, ak to vyzerá ako tá správna vec *aj po zohľadnení* problému skazeného hardvéru a katastrof čiernych labutí. Toto je rozumné, nie?

Existuje niekoľko odpovedí, ktoré na to mám.

Začnem tým, že toto je učebnicový príklad toho druhu myslenia, na ktoré som myslel, keď som varoval ašpirujúcich racionalistov, aby sa vyhýbali chytrosti.

Poznamenám aj, že by som nechcel, aby pokus o Priateľskú UI, ktorý sa práve rozhodol, že by sa Zem mala prerobiť na papierové spinky, posudzoval, či je to tá rozumná vec vo svetle všemožných varovaní, ktoré voči tomu dostal. Chcel by som, aby vykonal automatické kontrolované vypnutie. Kto tvrdí, že meta-uvažovanie bude odolné voči pokazeniu?

Mohol by som spomenúť dôležité situácie, v ktorých ma moje naivné, idealistické etické zábrany ochránili pred sebou samým, a postavili ma do pozície, z ktorej sa dalo zotaviť, alebo mi pomohli začať toto zotavovanie z veľmi hlbokých chýb, o ktorých som ani netušil, že ich robím. A mohol by som sa opýtať, či som naozaj pokročil tak ďaleko, a či by naozaj bolo také celkom múdre odstrániť ochrany, ktoré ma predtým zachránili.

Ale aj tak... „Som stále hlúpejší než moja etika?“ je otázka, na ktorú odpoveď nemusí byť *automaticky*: „Áno.“

Existujú samozrejmé hlúpe veci, ktoré by ste nemali robiť; napríklad by ste nemali čakať, dokiaľ nebudete v naozajstnom pokušení, a *potom* sa pokúšať zistiť, či ste chytrejší než vaša etika v tomto konkrétnom prípade.

Ale vo všeobecnosti... existuje len obmedzená sila, ktorú môžete udeliť tomu, čo vám vaši rodičia povedali, aby ste nerobili. Túto silu by ste nemali podceňovať. Chytrí ľudia debatovali o historických lekciách v procese vytvárania etiky Osvietenstva, z ktorej čerpá tak veľa západnej kultúry; a niektoré subkultúry ako je vedecká akadémia, alebo fandom vedeckej fantastiky, čerpajú z tejto etiky priamejšie. Ale aj tak je sila minulosti obmedzená.

A vlastne...

Ja som musel urobiť svoju etiku *omnoho prísnejšou* než čo mi moji rodičia a Jerry Pournelle a Richard Feynman povedali, aby som nerobil.

Je to zábavná vec, že keď sa ľuďom zdá, že sú chytrejší než ich etika, argumentujú za *menšiu* prísnosť, a nie za *väčšiu* prísnosť. Myslím tým, keď sa zamyslíte nad tým, o koľko zložitejší je moderný svet...

A v rovnakom duchu, tí ktorí prídu za mnou a povedia: „Mal by si klamať ohľadom Singularity, pretože tak dosiahneš, aby ťa podporovalo viac ľudí; je rozumné urobiť to, pre väčšie dobro“ -- títo zrejme nemajú *žiadnu predstavu* o rizikách.

Nespomínajú problém fungovania na skazenom hardvéri. Nespomínajú myšlienku, že lži musia byť rekurzívne chránené pre všetkými pravdami a pred všetkými technikami na hľadanie pravdy, ktoré ich ohrozujú. Nespomínajú, že poctivosť má svoju jednoduchosť, ktorá nečestným spôsobom často chýba. Nehovoria o stávkach na čierne labute. Nehovoria o hroznej nahote odhodenia poslednej obrany, ktorú máte proti sebe samému, a snahe prežiť na čistých výpočtoch.

Som silno presvedčený, že je to preto, lebo o týchto veciach *nemajú ani potuchy*.

Ak ste naozaj pochopili dôvod a rytmus za etikou, potom jeden z hlavných príznakov je to, že posilnení týmto novonadobudnutým poznaním *nerobíte* veci, ktoré vám predtým pripadali ako etické prehrešky. Teraz akurát už viete, prečo.

Niekto, kto sa iba pozrie na jeden alebo dva dôvody za etikou a povie: „Dobre, toto som pochopil, takže teraz to budem brať vedome do úvahy, a preto už nepotrebujem tieto etické zákazy“ -- taký sa správa viac ako stereotyp než ako skutočný racionalista. Svet nie je jednoduchý a čistý a jasný, takže nemôžete skrátka vziať etiku, s ktorou vás vychovali, a dôverovať jej. Ale toto predstieranie vulkánskej logiky, kde si myslíte, že len tak vypočítate všetko správne akonáhle ste dostali jeden alebo dva abstraktné vhľady -- ani toto v skutočnom živote nefunguje.

Pokiaľ ide o tých, ktorí na *nič* z tohto nedošli, a myslia si, že sú chytrejší než ich etika: Ha.

A čo sa týka tých, ktorí si predtým o sebe mysleli, že sú chytrejší než ich etika, ale ktorí si nepredstavovali všetky tieto prvky za etickými zákazmi „toľkými slovami“ dokiaľ nenatrafili na túto postupnosť na *Overcoming Bias*, a ktorí si myslia, že *teraz* už sú chytrejší než ich etika, pretože odteraz už budú toto všetko brať do úvahy: Dvakrát ha.

Videl som, ako sa mnoho ľudí usiluje vyhovoriť zo svojej etiky. Vždy je to zmena smerom k zhovievavosti, nikdy nie k väčšej prísnosti. A som ohúrený tou rýchlosťou a ľahkosťou, s akou sa snažia opustiť svoju ochranu. Hobbes povedal: „Neviem, čo je horšie, či skutočnosť, že každý má svoju cenu, alebo skutočnosť, že ich cena je taká nízka.“ Taká nízka cena, taký dychtiví sa nechať kúpiť. Nehľadajú alternatívy ešte po druhýkrát a po tretíkrát, skôr než sa rozhodnú, že im už nezostala žiadna iná možnosť než zhrešiť -- hoci sa môžu tváriť veľmi vážne a slávnostne, keď to hovoria. Opustia svoju etiku pri prvej príležitosti. „Kde existuje vôla zlyhať, tam sa nájdu prekážky.“ Vôľa zlyhať v etike sa u niektorých ľudí zdá byť veľmi silná.

Neviem, či môžem schvaľovať absolútne etické zákazy, záväzné pre všetky možné epistemické stavy ľudského mozgu. Vesmír nie je dosť láskavý na to, aby som mu v tomto dôveroval. (Hoci sa mi napríklad zdá, že etický zákaz sebaklamu má ohromnú silu. Videl som mnohých ľudí argumentovať za Temnú Stranu, a zdalo sa, že žiaden z nich si neuvedomuje sieťové riziká alebo riziká čiernych labutí pri sebaklame.) Ak sa jedného dňa pokúsim vytvoriť (reflexívne konzistentný) zákaz pre sebamodifikujúcu UI, bude to iba potom, čo vypracujem danú matematiku, pretože toto rozhodne nie je ten druh vecí, ktoré sa vám môžu prepiecť pomocou nejakej ad-hoc záplaty.

Ale poviem toľkoto:

Absolútne na mňa nerobia dojem vedomosti, uvažovanie, ani celková úroveň tých ľudí, ktorí za mnou dychtivo prídu a povedia vážnym hlasom: „Je rozumné urobiť neetickú vec X, pretože to bude mať výhodu Y.“

*→

289. Niečo chrániť

V prostredí (ehm) japonských rozprávok človek nachádza tento častý motív: Sila pochádza z toho, že je čo chrániť.

Nehovorím teraz iba o superhrdinoch, ktorí si pripnú zbrane, keď je priateľ ohrozený, ako to funguje v západných rozprávkach. V japonskej verzii to ide ešte hlbšie.

V ságe X je vyslovene uvedené, že každý kladný hrdina čerpá svoju silu z toho, že má niekoho – jedného človeka – ktorého chce chrániť. Koho? Táto otázka je časťou zápletky X – ten „najvzácnejší človek“ nie je vždy ten, kto si myslíme. Ale ak tohto človeka zabijú alebo mu škaredo ublížia, jeho ochranca stráca svoju moc – ani nie z nejakého magického spätného nárazu, ako z obyčajného zúfalstva. Toto nie je niečo, čo by sa každému kladnému hrdinovi stávalo každý týždeň, ako to býva v západných komixoch. Je to ekvivalentné zabitiu natrvalo – odstráneniu z hracej plochy.

V západných komixoch o superhrdinoch to funguje tak, že kladného hrdinu uhryzne rádioaktívny pavúk; a on potom musí niečo robiť so svojimi schopnosťami, nejako tráviť čas, tak sa rozhodne bojovať proti zločinu. Potom západní superhrdinovia vždy hundrú, koľko času im zaberajú ich superhrdinské povinnosti, a ako by radšej boli obyčajnými smrteľníkmi, aby mohli ísť na rybačku alebo také čosi.

Podobne v skutočnom západnom živote hovoríme nešťastným ľuďom, že potrebujú „životný cieľ“, takže by si mali vybrať nejaký altruistický prípad, ktorý sa hodí k ich osobnosti, napríklad vyberať vhodné bytové záclony, a toto ožiari ich dni pridaním ďalšej farby, tak ako keď máte pekné bytové záclony. Mali by ste si však dať pozor, aby ste si nevybrali niečo príliš drahé.

V západnom komixe prichádza najprv kúzlo, potom cieľ: Nadobudnete zázračné schopnosti, rozhodnete sa chrániť nevinných. V japonských rozprávkach to funguje naopak.

Samozrejme toto všetko nehovorím preto, aby som zovšeobecňoval fiktívne indície. Ale chcem sprostredkovať pojem, ktorého klamlivo podobná západná analógia *nie je* to, čo tým myslím.

Už som tu predtým naznačil myšlienku, že racionalista musí mať niečo, čo si cení viac než „rozumnosť“: *Umenie musí mať cieľ iný než samo seba, inak sa zrúti do nekonečnej rekurzie.* Ale nechápte ma zle, a nemyslite si, že odporúčam, aby si racionalisti vybrali nejaký pekný altruistický prípad, aby mali čo robiť, pretože rozumnosť samotná nie je až taká dôležitá. Nie. Ja sa pýtam: Odkiaľ racionalisti pochádzajú? Ako získavame svoje schopnosti?

Je napísané v Dvanástich cnostiach rozumnosti:

> Ako môžete zlepšiť svoje chápanie rozumnosti? Nie tak, že si poviete: „Je mojou povinnosťou byť rozumný.“ Tým by ste iba chránili svoje pomýlené predstavy. Možno je vaša predstava rozumnosti, že je rozumné veriť slovám Veľkého Učiteľa, a Veľký Učiteľ hovorí: „Obloha je zelená,“ a vy sa pozriete na oblohu a vidíte, že je modrá. Ak si myslíte: „Môže to vyzerať, že obloha je modrá, ale rozumnosť znamená veriť slovám Veľkého Učiteľa,“ stratili ste šancu odhaliť svoj omyl.

Historicky povedané, ľudstvo *konečne* opustilo pascu autority a začalo venovať pozornosť skutočnej oblohe tak, že názory založené na pokusoch sa ukázali ako *omnoho užitočnejšie* než názory založené na autorite. Zvedavosť tu bola od úsvitu ľudstva, ale problém je, že vymýšľanie rozprávok pri táboráku dokáže zvedavosť uspokojiť rovnako dobre.

Historicky povedané, veda vyhrala preto, lebo ukázala väčšiu hrubú silu vo forme technológie, nie preto, že by veda *znela rozumnejšie*. Do dnešného dňa kúzla a sväté písma stále znejú necvičeným ušiam rozumnejšie než veda. Preto tu stále trvá spoločenský konflikt medzi systémami viery. Keby veda nielen fungovala lepšie než kúzla, ale *aj* znela intuitívne rozumnejšie, už by bola dávno *úplne* vyhrala.

Sú aj takí, ktorí povedia: „Ako sa opovažuješ naznačiť, že niečo môže byť hodnotnejšie než Pravda? Vari nemusí racionalista milovať Pravdu viac než púhu užitočnosť?“

→ http://lesswrong.com/lw/v1/ethical_injunctions/

Zabudnime na chvíľku na to, čo by sa historicky stalo takémuto človeku – že ľudia s hodne podobnými postojmi mysle chránili Bibliu preto, lebo milovali Pravdu viac než púhu presnosť. Výroková morálka je slávna vec, ale má priveľa stupňov voľnosti.

Nie, skutočná pointa je, že vzťah lásky medzi racionalistom a Pravdou je, nuž, *ešte komplikovanejší* než emocionálny vzťah.

Človek sa nestane učňom rozumnosti bez toho, že by mu na pravde záležalo, aj ako na čisto morálnej hodnote, aj ako na niečom, čo je príjemné. Pochybujem, že existuje veľa majstrov skladateľov, ktorí nenávidia hudbu.

Ale časť toho, čo sa mi na rozumnosti *páči*, je disciplína ukladaná požiadavkou, aby názory dávali predpovede, čo nás dostáva k pravde omnoho bližšie než keby sme sedeli v obývačke a celý deň špekulovali o Pravde. *Páči* sa mi zložitosť toho, že treba zároveň milovať pravdivo vyzerajúce myšlienky, a pritom byť pripravený okamžite ich vyhodiť z okna. Páči sa mi aj slávna estetická čistota vyhlásenia, že si púhu užitočnosť vážim viac než estetiku. To je takmer rozpor, ale nie celkom; a aj toto má svoju estetickú stránku, podobne ako humor.

A samozrejme, nez ohľadu na to, ako veľmi vyznávate svoju lásku k púhej užitočnosti, nikdy by ste nemali *naozaj* skončiť tak, že budete naschvál veriť užitočnému nepravdivému výroku.

Nezjednodušujme si teda príliš tento vzťah medzi milovaním pravdy a milovaním užitočnosti. Nie je to jedno alebo druhé. Je to *zložité*, čo nemusí nutne byť defekt v morálnej estetike jednotlivých udalostí.

Avšak samotná morálka a estetika, viera, že človek by mal byť „rozumný“ alebo že určité spôsoby myslenia sú „krásne“, vás neprivedie do stredu Cesty. Nebolo by to dostalo ľudstvo z jamy autority.

V Kruhovom altruizme rozoberám túto dilemu: Ktorej z týchto možností by ste dali prednosť:

1. Zachrániť 400 životov, naisto
2. Zachrániť 500 životov s pravdepodobnosťou 90 %; nezachrániť žiaden život s pravdepodobnosťou 10 %.

Môžete byť v pokušení postaviť sa na tribúnu s výkrikom: „Ako sa opovažuješ hazardovať s ľudskými životmi?“ Dokonca aj keby ste vy osobne boli jedným z tých 500 – ale neviete, ktorým – mohli by ste byť v pokušení spoliehať sa na upokojujúci pocit istoty, pretože naše vlastné životy majú pre nás často menšiu hodnotu než dobrá intuícia.

Ale ak je vaša drahá dcéra jedna z tých 500, a vy neviete, *ktorá*, potom možno cítite silnejšie nutkanie držať hubu a počítať – všimnúť si, že máte šancu 80 % zachrániť ju v tom prvom prípade, a šancu 90 % zachrániť ju v tom druhom.

A áno, každý v tom dave je niekoho syn alebo dcéra. Čo teda naznačuje, že by sme si mali vybrať túto druhú možnosť aj ako altruisti, aj ako starostliví rodičia.

Týmto som nechcel naznačovať, že život jedného človeka je hodnotnejší než 499 ľudí. Chcel som tým povedať, že v stávke musí byť *viac* než váš vlastný život, kým človek konečne začne byť dosť zúfalý, aby sa uchýlil k matematike.

Čo ak veríte, že je „rozumné“ vybrať si istotu možnosti číslo 1? Veľa ľudí si myslí, že „rozumnosť“ znamená vyberanie si iba tých metód, ktoré naisto fungujú, a odmietanie všetkej neistoty. Ale dúfajme, že vám na živote vašej dcéry záleží viac ako na „rozumnosti“.

Zachráni vás pýcha vo svoju vlastnú cnosť racionalistu? Nie ak veríte, že je cnostné vybrať si istotu. O rozumnosti sa dokážete naučiť niečo nové iba ak vám na živote vašej dcéry záleží viac než na vašej pýche racionalistu.

Možno sa z tejto skúsenosti dokonca naučíte niečo nové o rozumnosti, ak ste dostatočne vyspelí vo svojom Umení, aby ste povedali: „Musel som mať nesprávnu predstavu o rozumnosti“ a nie: „Aha, ako mi rozumnosť dáva nesprávnu odpoveď!“

(Základným problémom v staní sa majstrom racionalistom je, že potrebujete pomerne veľa rozumnosti na to, aby ste naštartovali tento proces učenia.)

Je pre vás presvedčenie, že by ste mali byť rozumnejší, dôležitejšie než váš život? Pretože, ako som už predtým poznamenal, riskovať svoj život nie je v porovnaní s inými vecami až také strašidelné. Byť

osamelým hlasom nesúhlasu v dave, keď sa na vás každý výsmešne pozerá, je *omnoho* strašidelnejšie než púhe ohrozenie vášho života, podľa prejavených preferencií adolescentov, ktorí sa opijú na párty a potom šoférujú domov. Vyžaduje to niečo hrozne dôležité, aby ste boli ochotní opustiť stádo. Ohrozenie vášho života nemusí byť dosť.

Je vaša vôľa byť rozumným silnejšia než vaša *pýcha*? Môže to tak byť, ak vaša vôľa k rozumnosti pochádza z vašej hrdosti na svoj sebaobraz racionalistu? Môže pomáhať – *veľmi* pomáhať – mať sebaobraz, ktorý hovorí, že ste ten typ človeka, ktorý čelý tvrdej pravde. Je užitočné mať príliš veľa sebaúcty na to, aby človek sám sebe vedome klamal alebo odmietal čeliť indíciám. Ale môže prísť chvíľa, keď si musíte pripustiť, že ste robili rozumnosť celkom zle. Potom vaša hrdosť, váš sebaobraz racionalistu, môžu byť príliš tvrdou prekážkou.

Ak ste boli hrdí na to, že veríte, čo hovorí Veľký Učiteľ – aj keď to znie tvrdo, aj keď by ste radšej neverili – o to horkejšia môže byť pilulka priznania, že váš Veľký Učiteľ je podvodník a vaše vznešené sebaobetovanie bolo nanič.

Odkiaľ získate vôľu napredovať ďalej?

Keď sa obzriem na svoje osobné putovanie za rozumnosťou – nielen historické putovanie ľudstva – nuž, vyrástol som vo veľmi silnom presvedčení, že by som mal byť rozumný. Toto zo mňa urobilo nadpriemerného Tradičného Racionalistu na štýl Feynmana a Heinleina, a nič viac. Neposunulo ma to za hranice učenia, ktoré som dostal. Začal som rásť *ďalej* ako racionalista až keď som mal niečo hrozne dôležité, čo som potreboval urobiť. Niečo omnoho dôležitejšie než moja hrdosť racionalistu, viac než môj život.

Až keď sa oddáte úspechu viac než ľubovoľnej vašej obľúbenej technike rozumnosti, začnete oceňovať tieto slová Miyamota Musashiho:[264]

> „Môžete vyhrať pomocou dlhej zbrane, a predsa môžete vyhrať aj pomocou krátkej zbrane. V skratke, škola Way of the Ichi je duch vyhrávania, bez ohľadu na zbraň a jej veľkosť.“
>
> --Miyamoto Musashi, Kniha piatich kruhov

Nemýľte si toto s konkrétnym učením o rozumnosti. Je to popis ako sa *naučíte* Cestu, začínajúc od zúfalej túžby uspieť. Nikto nezvládne Cestu dokiaľ v stávke nie je viac než jeho život. Viac než jeho pohodlie, dokonca viac než jeho hrdosť.

Nemôžete si len tak vybrať nejaké Poslanie, pretože cítite, že potrebujete hobby. Hľadajte „dobré poslanie“ a vaša myseľ len doplní nejaké štandardné klišé. Naučte sa ako násobiť, a možno rozoznáte drasticky dôležité poslanie, keď ho uvidíte.

Ale *ak* máte takéto poslanie, potom je správne a primerané použiť rozumnosť v jeho službách.

Striktne podriadiť estetiku rozumnosti vyššiemu poslaniu je súčasťou estetiky rozumnosti. Mali by ste tejto estetike venovať pozornosť: Nikdy sa nestanete majstrom rozumnosti schopným vyhrať pomocou ľubovoľnej zbrane, pokiaľ neoceňujete krásu pre ňu samotnú.

*→

290. Kedy (ne)používať pravdepodobnosti

Možno niektorým čitateľom tohto blogu príde ako prekvapenie, že nie vždy odporúčam používanie pravdepodobností.

Alebo skôr, nie vždy odporúčam, aby ľudia, v snahe vyriešiť svoje problémy, skúšali *vymyslieť* slovné pravdepodobnosti, a potom použili zákony teórie pravdepodobnosti alebo teórie rozhodovania na tieto práve vymyslené čísla, a potom použili výsledok ako svoj konečný názor alebo rozhodnutie.

Zákony pravdepodobnosti sú zákony, nie odporúčania, ale často je skutočný Zákon pre nás ľudí príliš zložitý na vypočítanie. Ak P != NP a vesmír nemá žiaden zdroj exponenciálnej výpočtovej sily,

264 Musashi, *Book of Five Rings*.

→ http://lesswrong.com/lw/nb/something_to_protect/

potom existujú aktualizácie podľa indícií príliš zložité na to, aby ich spočítala hoci aj superinteligencia – napriek tomu, že by tieto pravdepodobnosti boli celkom dobre určené, keby sme si ich mohli dovoliť vypočítať.

Niekedy teda nepoužijete teóriu pravdepodobnosti. Najmä ak ste človek a váš mozog sa vyvinul s rozličnými užitočnými algoritmami na uvažovanie v neistote, ktoré *nezahŕňajú* slovné priradenie pravdepodobnosti.

Nie ste si istí, kam dopadne letiaca lopta? Neodporúčam vám, aby ste sa pokúsili sformulovať distribúciu pravdepodobnosti pre miesta jej dopadu, vykonali vedomé bayesovské aktualizácie podľa vášho pohľadu na loptu, a vypočítali očakávaný úžitok všetkých možných reťazcov pohybových pokynov pre svoje svaly. Ak sa pokúšate chytiť letiacu loptu, asi sa vám bude dariť lepšie s mechanizmami zabudovanými vo vašom mozgu, než používaním vedomého slovného uvažovania na vymýšľanie alebo manipulovanie pravdepodobností.

Ale to neznamená, že idete *mimo* teórie pravdepodobnosti alebo *nad* teóriu pravdepodobnosti.

Argumenty o dutch book stále platia. Ak vám ponúknem na výber stávky (10 000 dolárov, ak lopta dopadne v tomto štvorci, alebo 10 000 dolárov, ak hodím kockou a padne 6) a vy odpoviete spôsobom, ktorý nedovoľuje konzistentné priradenie pravdepodobností, potom budete prijímať kombinácie stávok, ktoré sú čistá strata, alebo odmietať stávky, ktoré sú čistý zisk...

Čo stále neznamená, že by ste sa mali pokúsiť použiť vedomé slovné uvažovanie. Očakával by som, že prinajmenšom pre profesionálnych hráčov baseballu je dôležitejšie chytiť loptu, než priradiť konzistentné pravdepodobnosti. Naozaj, keby ste sa pokúšali vymyslieť pravdepodobnosti, *slovné* pravdepodobnosti by ani nemuseli byť veľmi dobré, v porovnaní s nejakým inštinktívnym pocitom – nejakou bezslovnou reprezentáciou neistoty v pozadí vašej mysle.

Nie je nič privilegované na neistote, ktorá je vyjadrené slovami, jedine ak slovné časti vášho mozgu *naozaj* fungujú na daný problém lepšie.

A hoci presné mapy toho istého územia budú nevyhnutne navzájom konzistentné, nie všetky konzistentné mapy sú presné. Je dôležitejšie byť presný než konzistentný, a dôležitejšie chytiť loptu než byť konzistentný.

V skutočnosti vo všeobecnosti odporúčam *nevymýšľať si* pravdepodobnosti, pokiaľ to nevyzerá, že máte na to nejaký slušný základ. Iba by vás to zmiatlo veriť, že ste bayesovskejší než naozaj ste.

Konkrétnejšie by som vo väčšine prípadov odporúčal nepoužívať nečíselné procedúry na vytvorenie niečoho, čo vyzerá ako číselná pravdepodobnosť. Čísla by mali pochádzať z čísel.

Áno, *existujú* výhody, keď sa pokúšate preložiť svoje inštinktívne pocity neistoty do slovných pravdepodobností. Môže vám to pomôcť zbadať problémy ako je klam konjunkcie. Môže vám to pomôcť zbadať vnútorné nekonzistencie – hoci vám to nemusí ukázať žiaden spôsob, ako ich vyriešiť.

Ale nemali by ste chodiť s predstavou, že ak preložíte svoj inštinktívny pocit ako „jeden z tisíca“, potom v prípadoch, keď vyslovíte tieto slová, sa zodpovedajúca udalosť stane približne v jednom prípade zo tisíca. Váš mozog nie je tak dobre kalibrovaný. Ak namiesto toho urobíte niečo neverbálne so svojím inštinktívnym pocitom neistoty, možno na tom budete lepšie, pretože aspoň *použijete* ten inštinkt tak, ako bol zamýšľaný.

Táto konkrétna téma sa nedávno vynorila v kontexte Veľkého hadronového kolidera, a debaty na konferencii o rizikách globálnej katastrofy:

Že si nemôžeme byť istí, že nie je žiadna chyba v článkoch, ktoré ukazovali z viacerých uhlov, že LHC nemôže zničiť celý svet. A navyše, teória použitá v tých článkoch môže byť nesprávna. A v oboch prípadoch je stále šanca, že LHC *môže* zničiť celý svet. A preto by sa nemal zapínať.

Nuž, keby ten argument bol postavený *iba* takto, nenamietal by som voči jeho epistemológii.

Ale rečník sa v skutočnosti pokúšal priradiť pravdepodobnosť aspoň 1 z 1000, že teória, model alebo výpočty v článkoch o LHC sú nesprávne; a pravdepodobnosť aspoň 1 z 1000, že ak sú teória, model alebo výpočty nesprávne, LHC zničí celý svet.

Napokon, iste nie je také nepravdepodobné, že budúce generácie odmietnu teóriu použitú v článku o LHC, alebo odmietnu model, alebo možno len nájdu chybu. A ak je článok o LHC nesprávny, potom kto vie, čo sa v dôsledku toho môže stať?

Takže toto je argument – ale priradiť mu *čísla*?

Namietam proti dojmu autority vytvorenému tým, že sa tieto čísla vycucali z prsta. Vo všeobecnosti mám pocit, že ak nemôžete použiť pravdepodobnostné nástroje na tvarovanie svojich pocitov neistoty, nemali by ste ich poctiť tým, že ich nazvete pravdepodobnosti.

Alternatíva, ktorú by som v tomto konkrétnom prípade navrhol, je debatovať o všeobecnom pravidle zakázania fyzikálnych pokusov, pretože si nemôžete byť absolútne istí argumentami, ktoré hovoria, že sú bezpečné.

Trvám na tom, že keby ste to formulovali takto, potom si vaša myseľ po zvážení frekvencie udalostí pravdepodobne prinesie do rozhodovania viac dôsledkov, a spomenie si na viac relevantných historických prípadov.

Ak debatujete iba o jednom prípade LHC a priradíte konkrétne pravdepodobnosti, (1) dáva to veľmi chabému uvažovaniu neprimeraný dojem autority, (2) skrýva to všeobecné dôsledky aplikovania podobných pravidiel, a dokonca (3) vytvára ilúziu, že by sme mohli prísť k inému rozhodnutiu, keby niekto iný uverejnil nový fyzikálny článok, ktorý by znížil tieto pravdepodobnosti.

Zdalo sa, že autori na konferencii o riziku globálnej katastrofy naznačujú, že by sme mohli len urobiť trochu viac analýzy LHC a *potom* ho zapnúť. To mi prišlo ako tá nejneúprimnejšia časť argumentu. Akonáhle pripustíte argument: „Možno je analýza zlá a ktovie, čo sa stane potom," potom neexistuje fyzikálny článok, ktorý by sa ho kedy mohol zbaviť.

Bez ohľadu na to, aké iné fyzikálne články boli publikované predtým, autori by použili ten istý argument a *vymysleli by si tie isté číselné pravdepodobnosti* na konferencii o riziku globálnej katastrofy. Nemôžem si byť touto vetou istý, samozrejme, ale má pravdepodobnosť 75 %.

Vo všeobecnosti sa racionalista pokúša dosiahnuť, aby jeho myseľ fungovala s najväčším dosiahnuteľným výkonom; to niekedy zahŕňa *hovorenie* o slovných pravdepodobnostiach, a niekedy nie, ale zákony pravdepodobnosti *platia* vždy.

Ak všetko, čo máte, je inštinktívny pocit neistoty, potom by ste sa pravdepodobne mali držať tých algoritmov, ktoré používajú inštinktívny pocit neistoty, pretože vaše zabudované algoritmy môžu fungovať lepšie než vaše nešikovné pokusy vyjadriť veci slovami.

Je celkom možné, že pri takomto uvažovaní zistím, že som nekonzistentný. Napríklad, podstatne viac by ma vydesil žrebovací stroj, ktorý by mal presne definovanú šancu 1 z 1 000 000, že zničí celý svet, než zapnutie Veľkého hadronového kolidera.

Na druhej strane, keby ste sa ma opýtali, či by som vedel urobiť jeden milión viet s rovnakou autoritou ako „Veľký hadrónový kolider nezničí celý svet" a mýliť sa v priemere jedenkrát, potom by som musel povedať nie.

Čo by som mal robiť s touto nekonzistenciou? Nie som si istý, ale určite nebudem mávať čarovnou paličkou, aby zmizla. To je ako nájsť nekonzistenciu medzi dvoma mapami, ktoré máte, a rýchlo načmárať nejaké zmeny, aby sa dosiahla ich konzistentnosť.

Tiež by som sa, mimochodom, podstatne viac obával žrebovacieho stroja s pravdepodobnosťou 1 k 1 000 000 000 zničenia celého sveta než stroja, ktorý zničí celý svet, ak existuje židovsko-kresťanský Boh. Ale nepredpokladám, že by som dokázal urobiť milión tvrdení, jedno za druhým, celkom nezávislé a rovnako vážne ako „Neexistuje žiaden Boh" a mýliť sa v priemere jedenkrát.

Nemôžem povedať, že som *spokojný* s takýmto stavom epistemických záležitostí, ale nebudem ho meniť, dokiaľ neuvidím, že sa hýbem smerom k *väčšej presnosti a efektivite v skutočnom svete*, nie iba smerom k väčšej vnútornej konzistencii. Napokon, cieľom je vyhrať. Ak si vymyslím pravdepodobnosť, ktorá nie je tvarovaná pravdepodobnostnými nástrojmi, ak si vymyslím číslo, ktoré nebolo vytvorené numerickými metódami, potom možno iba sabotujem svoj zabudovaný mechanizmus, ktorý by urobil lepšie, keby fungoval v svojom prirodzenom režime neistoty.

Toto samozrejme nie je povolenie ignorovať pravdepodobnosti, ktoré sú dobre podložené. Hocijaký numerický základ je pravdepodobne lepší než nejasný pocit neistoty; ľudia sú mizerní štatistici. Ale vycucať si číslo *úplne* z prsta, čiže použiť nenumerickú procedúru na vytvorenie čísla, to nie je takmer žiaden základ; a v takom prípade ste na tom asi lepšie, keď zostanete s nejasným pocitom neistoty.

Čo je dôvod, prečo moje články na Overcoming Bias zvyčajne používajú slová ako „možno“ a „pravdepodobne“ a „určite“ namiesto priraďovania vymyslených číselných pravdepodobností ako „40 %“ a „70 %“ a „95 %“. Predstavte si, ako hlúpo by to vyzeralo. Myslím si, že by to naozaj *bolo* hlúpe; myslím si, že by to tak bolo horšie.

Nie som ten druh slameného bayesiána, ktorí hovorí, že by ste si mali vymyslieť pravdepodobnosti, aby ste sa vyhli možným dutch books. Som ten druh bayesiána, ktorí hovorí, že v praxi sa ľudia stanú obeťou dutch books, pretože nie sú dosť silní na to, aby sa im vyhli; a navyše je dôležitejšie chytiť loptu než vyhnúť sa dutch books. Tá matematika je ako *základná* fyzika, niet úniku pred jej vládou, ale je príliš drahé ju počítať.

Nemá zmysel žiaden poznávací rituál, ktorý *napodobňuje* povrchnú podobu matematiky, ale nevie produkovať systematicky lepšie rozhodovanie. To by bol stratený účel; to nie je to pravé umenie života podľa zákona.

*→

291. Newcombov problém a ľutovanie rozumnosti

Nasledujúca úloha je asi tá najkontroverznejšia dilema v histórii teórie rozhodovania:

> Superinteligencia z inej galaxie, ktorú nazývame Omega, príde na Zem a začne hrať čudnú hru. V tejto hre si Omega vyberie nejakého človeka, postaví pred neho dve krabice a odletí preč.
>
> Krabica A je priehľadná a obsahuje tisíc dolárov.
>
> Krabica B je nepriehľadná a obsahuje buď milión dolárov alebo nič.
>
> Môžete si vziať obe krabice, alebo vziať iba krabicu B.
>
> A chyták je v tom, že Omega dala do krabice B milión dolárov vtedy a iba vtedy, ak Omega predpovedala, že si vezmete iba krabicu B.
>
> Omega mala pravdu v každom z doteraz pozorovaných 100 prípadov – každý, kto si vzal obe krabice, zistil, že krabica B je prázdna a dostal iba tisíc dolárov; každý, kto si vzal iba krabicu B, zistil, že krabica B obsahuje milión dolárov. (Predpokladáme, že ak si vezmete iba krabicu B, krabica A sa rozplynie v oblaku dymu; nikto iný si nemôže dodatočne vziať krabicu A.)
>
> Skôr než si vyberiete, Omega už odletela a hrá niekde inde ďalšiu svoju hru. Krabica B už je prázdna, alebo už je plná.
>
> Omega položí dve krabice na zem pred vás a odletí preč.
>
> Vezmete si obe krabice, alebo iba krabicu B?

A štandardná filozofická konverzácia sa vyvíja takto:

> Jednokrabičkár: „Ja si samozrejme vezmem iba krabicu B. Radšej budem mať milión ako tisíc.“
>
> Dvojkrabičkár: „Ale Omega už odišla. Buď je krabica B už plná alebo je už prázdna. Ak je krabica B už prázdna, potom vziať obe krabice mi dá 1000 dolárov, a vziať iba krabicu B mi dá 0 dolárov. Ak je krabica B už plná, potom vziať obe krabice mi dá 1 001 000 dolárov, a

→ http://lesswrong.com/lw/sg/when_not_to_use_probabilities/

vziať iba krabicu B mi dá 1 000 000 dolárov. V oboch prípadoch je výhodnejšie vziať obe krabice, a menej výhodné nechať tisíc dolárov ležať – budem teda rozumný a vezmem si obe krabice."

Jednokrabičkár: „Keď si taký rozumný, prečo nie si aj bohatý?"

Dvojkrabičkár: „Nie je to moja chyba, že sa Omega rozhodla odmeňovať iba ľudí s nerozumnými sklonmi, ale pre mňa je už neskoro, aby som s tým niečo urobil."

Na tému problémov ako je Newcombov existuje *rozsiahla* literatúra – najmä ak vezmeme Väzenskú dilemu ako jej špeciálny prípad, čo sa tak zvyčajne berie. *Paradoxy rozumnosti a spolupráce*[265] je editovaný zborník, ktorý obsahuje pôvodný Newcombov článok. Pre tých, ktorí čítajú iba materiály na internete, Ledwigova dizertačná práca zhŕňa najčastejšie štandardné postoje.[266]

Nebudem tu preberať celú literatúru, ale v modernej teórii rozhodovania prevláda zhoda, že človek by mal vziať obe krabice, a že Omega jednoducho odmeňuje činiteľov s nerozumnými sklonmi. Tento prevládajúci názor sa nazýva „kauzálna teória rozhodovania".

Nebudem tu skúšať predkladať svoju vlastnú analýzu. To je príliš dlhý príbeh, ešte aj na moje pomery.

Ale dokonca aj kauzálni rozhodovací teoretici sa zhodnú, že ak máte schoposť vopred sa zaviazať, že v Newcombovom probléme vezmete iba jednu krabicu, mali by ste to urobiť. Ak sa tak zaviažete ešte predtým ako vás Omega preskúma, potom ste priamo spôsobili, že krabica B bude plná.

Moja oblasť – v prípade, že ste zabudli, je to sebamodifikujúca UI – toto v preklade znamená, že ak postavíte UI, ktorá by si v Newcombovom probléme vybrala dve krabice, táto sa zmení, aby si v Newcombovom probléme vyberala iba jednu krabicu, ak predpokladá, že sa dostane do takejto situácie. Činitelia s voľným prístupom k svojmu zdrojovému kódu majú lacnú metódu zaväzovania sa.

Čo ak očakávate, že by ste sa vo všeobecnosti mohli stretnúť s problémom ako je ten Newcombov, ale nepoznáte presnú podobu problému? Potom by ste sa mohli zmeniť na taký druh činiteľa, ktorý má také sklony, že vo všeobecnosti dostáva vysoké odmeny v problémoch ako je ten Newcombov.

Ako ako vlastne vyzerá taký činiteľ, ktorý je vo všeobecnosti dobre prispôsobený problémom ako je ten Newcombov? Dá sa to formálne špecifikovať?

Áno, ale keď som to skúšal napísať, uvedomil som si, že začínam písať malú knihu. A nebola to tá najdôležitejšia kniha na napísanie, tak som to odložil do šuflíka. Moje pomalé písanie je naozaj kliatbou mojej existencie. Zdá sa, že teória, na ktorej som pracoval, má veľa pekných vlastností okrem toho, že sa hodí na problémy ako je ten Newcombov. Bola by z nej pekná dizertačná práca, keby som našiel niekoho, kto by to prijal ako moju dizertačnú prácu. Ale to je viacmenej jediný dôvod, prečo by som tento projekt opäť vybral zo šuflíka. Inak neviem zdôvodniť taký časový výdavok, nie pri rýchlosti, akou teraz píšem knihy.

Toto všetko hovorím, pretože existuje častý postoj, že „Slovné argumenty pre jednu krabicu sa ľahko vymýšľajú, ťažké je však vyvinúť dobrú teóriu rozhodovania, ktorá si vyberá jednu krabicu" - koherentnú matematiku, ktorá si vyberá jednu krabicu v Newcombovom probléme bez toho, že by niekde inde produkovala absurdné výsledky. Ja tomu rozumiem, aj som sa pokúsil vyvinúť takú teóriu, ale moje písanie veľkých článkov je také pomalé, že to nedokážem publikovať. Či mi veríte alebo nie, je to pravda.

Tak či tak by som rád predložil niektoré z mojich *motívov* ohľadom Newcombovho problému – dôvody, ktoré ma poháňajú hľadať novú teóriu – pretože ilustrujú moje základné postoje k rozumnosti. Aj keď neviem predložiť celú teóriu, ktorú tieto motivácie motivujú...

V prvom rade, ako základ, nado všetko ostatné:

Rozumní činitelia by mali VYHRÁVAŤ.

Nechápte ma zle, a nemyslite si, že hovorím o Hollywoodskom stereotype rozumnosti, že racionalisti by mali byť sebeckí alebo krátkozrakí. Ak vaša funkcia úžitku obsahuje člen pre druhých,

265 Richmond Campbell and Lanning Snowden, eds., *Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem* (Vancouver: University of British Columbia Press, 1985).

266 Marion Ledwig, „Newcomb's Problem" (PhD diss., University of Constance, 2000).

potom vyhrajte ich šťastie. Ak vaša funkcia úžitku obsahuje člen pre milión rokov vzdialenú budúcnosť, potom vyhrajte tento eón.

Ale v každom prípade *VYHRAJTE*. Neprehrajte rozumne, **VYHRAJTE**.

Existujú obrancovia kauzálnej teórie rozhodovania, ktorí hovoria, že dvojkrabičkári robia to najlepšie, čo môžu, aby vyhrali, a nemôžu si pomôcť, ak boli prekliati Predpovedačom, ktorý nadŕža nerozumným. Tejto obrane sa budem o chvíľu venovať. Ale najprv chcem ukázať rozdiel medzi kauzálnymi teoretikmi rozhodovania, ktorí veria, že dvojkrabičkári naozaj robia to najlepšie, čo môžu, aby vyhrali; verzus niekto, kto si myslí, že dvojkrabičkovanie je tá *rozumná* alebo *racionálna* vec, ktorú treba urobiť, ale že v tomto prípade tento rozumný krok zhodou okolností predvídateľne prehrá. Existuje *veľa* ľudí, ktorí si myslia, že rozumnosť predvídateľne prehráva v rôznych problémoch – aj to je súčasť hollywoodskeho stereotypu rozumnosti, kde je Kirk predvídateľne lepší než Spock.

Ďalej sa venujme obvineniu, že Omega nadŕža nerozumným. Viem si predstaviť superbytosť, ktorá odmeňuje iba ľudí narodených s konkrétnym génom, *bez ohľadu na ich voľby*. Viem si predstaviť superbytosť, ktorá odmeňuje ľudí, ktorých mozgy obsahujú *konkrétny algoritmus*: „Opíš svoje možnosti po anglicky, a potom si vyber poslednú možnosť v abecednom poradí," ale neodmeňuje nikoho, kto si vyberie rovnakú možnosť, ale z iného dôvodu. Lenže Omega odmeňuje ľudí, ktorí si vezmú iba krabicu B, *bez ohľadu na to, aký algoritmom došli k tomuto záveru*, a preto nesúhlasím s obvinením, že Omega odmeňuje nerozumných. Omega sa nestará o to, či ste sa riadili alebo neriadili nejakým konkrétnym rituálom poznávania; Omega sa stará iba o vaše predpovedané *rozhodnutie*.

Môžeme si vyberať ľubovoľným algoritmom uvažovania, a budeme odmenení alebo potrestaní iba za to, čo si tento algoritmus vyberie, na ničom inom to nezávisí – Omega sa stará o to, kam ideme, ale nie, ako sme sa tam dostali.

A práve toto pochopenie, že príroda sa nestará o nás *algoritmus*, nám umožňuje nasledovať Cestu vyhrávania – bez pripútanosti s ľubovoľnému konkrétnemu rituálu poznávania, okrem nášho názoru, že vyhráva. Každé pravidlo sa dá obísť, *okrem* pravidla, že vyhrávame.

Ako povedal Miyamoto Musashi – naozaj sa to oplatí zopakovať:

> Môžete vyhrať pomocou dlhej zbrane, a predsa môžete vyhrať aj pomocou krátkej zbrane. V skratke, škola Way of the Ichi je duch vyhrávania, bez ohľadu na zbraň a jej veľkosť.[267]

(Iný príklad: McGee tvrdil, že si musíme osvojiť obmedzené funkcie úžitku, inak sa môžeme dostať do nekonečných postupností „Dutch books". Lenže: *Funkcia úžitku sa nedá obchádzať.* Ja naozaj milujem život bez obmedzenia alebo hornej hranice: Neexistuje také konečné množstvo života N, že by som uprednostnil pravdepodobnosť 80,000 1 %, že budem žiť N rokov voči pravdepodobnosti 0,000 1 %, že budem žiť $10^{10^{100}}$ rokov a pravdepodobnosti 80 %, že budem žiť navždy. Toto je dostatočná podmienka na to, aby moja funkcia úžitku bola neohraničená. Musím teda prísť na to, ako optimalizovať *pre túto morálku*. Nemôžete mi najprv povedať, že sa musím predovšetkým pridŕžať nejakého konkrétneho rituálu poznávania, a potom, že ak sa pridŕžam tohto rituálu, musím zmeniť svoju morálku, aby som sa vyhol Dutch-bookovaniu. Zahoďte ten prehrávajúci rituál; nemeňte definíciu vyhrávania. To je ako keby sme sa rozhodli, že dávame prednosť 1000 dolárom pred 1 000 000 dolármi, aby Newcombov problém nestaval náš obľúbený rituál rozhodovania do zlého svetla.)

„Lenže," povie kauzálny teoretik rozhodovania, „aby ste vzali iba jednu krabicu, musíte nejako veriť, že vaša voľba dokáže ovplyvniť, či je krabica B prázdna alebo plná – a to je *nerozumné*! Omega už odišla! Je to fyzicky nemožné!"

Nerozumné? Ja som racionalista: prečo by som sa staral o to, čo je nerozumné? Nemusím sa podriaďovať žiadnemu konkrétnemu rituálu poznávania. Nemusím si vybrať krabicu B *iba kvôli tomu, že verím, že moja voľba ovplyvňuje krabicu, hoci Omega už odišla.* Môžem si jednoducho... zobrať iba krabicu B.

267 Musashi, *Book of Five Rings*.

Mám návrh alternatívneho rituálu poznávania, ktorý vypočíta tento výsledok, ale nezmestí sa na tento malý okraj; ale nemal by som vám ho potrebovať ukazovať. Pointa nie je mať elegantnú teóriu vyhrávania – pointa je vyhrať; elegancia je vedľajší účinok.

Alebo sa na to pozrime z inej strany: Namiesto aby sme začali od predstavy o tom, čo je rozumné rozhodovanie, a potom sa pýtali, či „rozumní“ činitelia odchádzajú s hromadou peňazí, začnime pozeraním sa na činiteľov, ktorí odchádzajú s hromadou peňazí, vyviňme si teóriu o tom, ktorí agenti zvyknú odchádzať s najväčším množstvom peňazí, a z tejto teórie sa pokúsme zistiť, čo je „rozumné“. Slovo „rozumné“ možno odkazuje iba na rozhodnutia, ktoré sú v súlade s naším súčasným rituálom poznávania – čo iné by ešte mohlo určovať, či niečo vyzerá „rozumne“ alebo nie?

Od Jamesa Joycea, *Základy kauzálnej teórie rozhodovania*:[268]

> Rachel má úplne dobrú odpoveď na otázku: „Prečo nie si bohatá?“ „Nie som bohatá,“ povie, „pretože nie som ten typ človeka, o ktorom si psychológ myslí, že by odmietol tieto peniaze. Skrátka nie som ako ty, Irene. Keďže viem, že som ten typ, ktorý si vezme peniaze, a keďže ten psychológ vie, že som tento typ, je rozumné, aby som si myslela, že tých 1 000 000 dolárov nie je na mojom účte. Tých 1000 dolárov je maximum, čo môžem získať, nech robím čokoľvek. Takže jediná rozumná možnosť pre mňa je vziať si ich.“
>
> Irene môže chcieť ešte trochu pritlačiť a opýta sa: „Ale neželala by si, aby si bola ako ja, Rachel? Neželáš si, aby si bola ten odmietajúci typ?“ Máme sklon si myslieť, že Rachel, ako oddaná kauzálna teoretička rozhodovania, musí na túto otázku odpovedať záporne, čo znie očividne nesprávne (keďže keby bola ako Irene, bola by bohatá). Toto nie je ten prípad. Rachel môže a mala by priznať, že by radšej bola taká ako Irene. „Bolo by pre mňa lepšie,“ môže priznať, „keby som bola ten odmietajúci typ.“ V tomto bode Irene vyhlási: „Priznala si to! Nebolo teda až také chytré vziať si tie peniaze.“ Nanešťastie pre Irene, jej záver nevyplýva z Racheliných predpokladov. Rachel jej trpezlivo vysvetlí, že želať si byť odmietačom v Newcombovom probléme nie je nekonzistentné s myslením si, že človek by mal vziať tých 1000 dolárov *bez ohľadu na to, aký typ je*. Keď si Rachel želá, aby bola ten typ ako Irene, želá si, aby *mala také možnosti* ako Irene, neschvaľuje tým jej voľbu.

Povedal by som, že je všeobecným princípom rozumnosti – ba priam časť toho, ako rozumnosť *definujem* – že by ste nikdy nemali skončiť tak, že niekomu závidíte samotnú jeho *voľbu*. Môžete niekomu závidieť jeho gény, ak Omega odmeňuje gény, alebo ak mu gény vo všeobecnosti dávajú šťastnejší život. Ale v tomto prípade Rachel závidí Irene jej voľbu, a *iba* jej voľbu, bez ohľadu na to, aký algoritmus Irene použila, aby k nej došla. Rachel si želá *iba*, aby mala schopnosť vybrať si inak.

Nemali by ste tvrdiť, že ste rozumnejší než niekto a zároveň mu závidieť jeho voľbu – *iba* voľbu samotnú. Jednoducho *konajte* tým spôsobom, ktorý závidíte.

Pokúšam sa povedať, že rozumnosť je Cesta vyhrávania, ale kauzálni rozhodovací teoretici trvajú na tom, že vziať si obe krabice je to, čo *naozaj* vyhráva, pretože *predsa nemôže byť lepšie* nechať 1000 dolárov na stole... aj keď jednokrabičkári odchádzajú z pokusu s väčším množstvom peňazí. Dávajte si pozor na tento typ argumentu, vždy keď sa pristihnete, že definujete ako „víťaza“ niekoho iného než toho činiteľa, ktorý sa práve usmieva z vrchu vysokej hromady úžitku.

Áno, existujú rôzne myšlienkové pokusy, v ktorých niektorí činitelia začínajú s výhodou – ale ak je úlohou povedzme rozhodnúť sa, či skočíte z útesu, dávajte si pozor, aby ste nedefinovali činiteľov, ktorí neskočia z útesu ako tých, čo majú neférovú výhodu oproti tým, ktorí skočia z útesu, len preto že odmietajú skočiť z útesu. V tomto bode ste nenápadne predefinovali „víťazstvo“ na dodržiavanie nejakého konkrétneho rituálu poznávania. *Sledujte peniaze!*

Alebo tu je iný spôsob, ako na to pozerať: Keby ste sa stretli s Newcombovým problémom, chceli by ste naozaj pozorne hľadať dôvod veriť, že je dokonale rozumné a racionálne vziať si iba krabicu B; pretože, keby taký argument existoval, mohli by ste si vziať iba krabicu B a zistiť, že plná peňazí? Boli by

268 James M. Joyce, *The Foundations of Causal Decision Theory* (New York: Cambridge University Press, 1999), doi:10.1017/CBO9780511498497.

ste ochotní venovať tomu hodinu navyše, aby ste si to premysleli, keby ste si boli istí, že na konci tejto hodiny dokážete presvedčiť sami seba, že krabica B je tá rozumná voľba? Aj toto je pomerne nezvyčajná situácia. Za bežných okolností je úlohou rozumnosti zistiť, ktorá voľba je najlepšia – nie nájsť dôvod, prečo veriť, že nejaká konkrétna možnosť je najlepšia.

Možno je príliš ľahké povedať, že „by ste mali“ vziať dve krabice v Newcombovom probléme, že toto je tá „rozumná“ vec, pokiaľ tie peniaze nie sú naozaj pred vami. Možno ste dovtedy len boli unavení filozofickými dilemami. Čo keby vaša dcéra mala chorobu, ktorá je na 90 % smrteľná, a v krabici A by bolo sérum, ktoré ju uzdraví s pravdepodobnosťou 20 %, a v krabici B by bolo sérum, ktoré ju uzdraví s pravdepodobnosťou 95 %? Čo keby sa na Zem rútil asteroid, a krabica A by obsahovala odchyľovač asteroidov, ktorý funguje v 10 % prípadov, a krabica B by obsahovala odchyľovač asteroidov, ktorý funguje v 100 % prípadov?

Boli by ste v takomto momente *v pokušení vybrať si nerozumne?*

Keby v krabici B bolo niečo, čo *nemôžete* nechať len tak? Niečoho omnoho dôležitejšie než to, či sa správate rozumne? Keby ste absolútne *museli* vyhrať – *naozaj* vyhrať, nie len definovať si, že ste vyhrali?

Želali by ste si zo všetkých síl, aby „rozumné“ rozhodnutie bolo vziať si iba krabicu B?

Potom je možno načase aktualizovať svoju definíciu rozumnosti.

Údajní racionalisti by nemali zisťovať, že závidia púhe rozhodnutia údajných neracionalistov, pretože vaše rozhodnutie môže byť také, aké len chcete. Keď zistíte, že ste v takejto pozícii, nemali by ste karhať toho druhého za to, že sa nepodriaďuje vašej predstave rozumnosti. Mali by ste si uvedomiť, že chápete Cestu nesprávne.

Podobne je to, keby ste sa niekedy pristihli, že si osobitne pamätáte „rozumné“ názory, verzus názory, ktoré sú pravdepodobne naozaj *pravdivé*. Buď ste nepochopili rozumnosť, alebo je vaša druhá intuícia jednoducho nesprávna.

Nemôžete však zároveň *definovať* „rozumnosť“ ako vyhrávajúcu Cestu, a *definovať* „rozumnosť“ ako bayesovskú teóriu pravdepodobnosti a teóriu rozhodovania. Ale to je ten argument, ktorý tu predkladám, a ponaučenie mojej rady Dôverovať Bayesovi, že zákony, ktorými sa riadi vyhrávanie, sa naozaj ukázali byť matematika. Keby sa niekedy ukázalo, že Bayes zlyháva – dostáva systematicky nižšie odmeny za niektoré problémy, v porovnaní s lepšou alternatívou, na základe púhych svojich odpovedí – potom Bayesa treba *vyhodiť z okna*. „Rozumnosť“ je iba nálepka, ktorou označujem moje názory na vyhrávajúcu Cestu – na Cestu činiteľa, ktorý sa usmieva z vrchu obrovskej hromady úžitku. *V súčasnosti* táto nálepka označuje bayesovské remeslo.

Uvedomujem si, že toto nie je zdrvujúca kritika kauzálnej teórie rozhodovania – to by si vyžiadalo skutočnú knihu a/alebo dizertačnú prácu – ale dúfam, že to ilustruje niektoré moje základné postoje k tomuto poňatiu „rozumnosti“.

[Doplnené 2015: Už som napísal knihu s úvodom do rozhodovacej teórie, ktorá dominuje kauzálnej teórii rozhodovania, „Bezčasová teória rozhodovania“.[269] Kryptograf Wei Dai reagoval inou alternatívou ku kauzálnej teórii rozhodovania, bezaktualizačnou teóriou rozhodovania, ktorá dominuje aj kauzálnej aj bezčasovej teórii rozhodovania. V roku 2015 sú najlepšími aktuálnymi diskusiami k týmto teóriám Daniel Hintze: „Dominancia tried problémov v predpovedacích dilemách“[270] a Nate Soares a Benja Fallenstein: „Smerom k idealizovanej teórii rozhodovania“.[271]]

Nemali by ste zistiť, že rozlišujete vyhrávajúcu voľbu od rozumnej voľby. Nemali by ste zistiť, že rozlišujete rozumný názor od názoru, ktorý je s najväčšou pravdepodobnosťou pravdivý.

To je dôvod, prečo používam slovo „rozumný“ na označenie mojich názorov na presnosť a vyhrávanie – *nie* aby som označil slovné uvažovanie, alebo stratégie, ktoré prinášajú zaručený úspech, alebo to, čo sa dá logicky dokázať, alebo to, čo sa dá verejne predviesť, alebo to, čo je rozumné.

Ako povedal Miyamoto Musashi:

269 Yudkowsky, *Timeless Decision Theory*.

270 Daniel Hintze, „Problem Class Dominance in Predictive Dilemmas,“ Honors thesis (2014).

271 Nate Soares and Benja Fallenstein, „Toward Idealized Decision Theory,“ Technical report. Berkeley, CA: Machine Intelligence Research Institute (2014), http://intelligence.org/files/TowardIdealizedDecisionTheory.pdf.

Prvoradou vecou, keď uchopíš meč do rúk, je tvoj zámer ťať nepriateľa, akýmkoľvek spôsobom. Kedykoľvek odrážaš, udieraš, bodáš, sekáš, alebo sa dotýkaš nepriateľovho tnúceho meča, musíš v tom istom pohybe ťať nepriateľa. Je podstatné toto dosiahnuť. Ak myslíš iba na odrážanie, udieranie, bodanie, sekanie, alebo dotýkanie sa nepriateľa, nedokážeš ho naozaj sťať.

*→

Medzihra: Dvanásť cností rozumnosti

Prvou cnosťou je zvedavosť. Pálčivé nutkanie vedieť znamená viac ako slávnostná prísaha nasledovať pravdu. Cítiť pálčivé nutkanie zvedavosti vyžaduje, aby si bol nevedomý, a aby si zároveň túžil zanechať svoju nevedomosť. Ak v hĺbke srdca veríš, že už vieš, alebo ak si v hĺbke srdca neželáš vedieť, potom tvoje vypytovanie bude bez cieľa a tvoje zručnosti bez účelu. Zvedavosť sa snaží zničiť sama seba; niet takej zvedavosti, ktorá by nechcela odpoveď. Krásou krásneho tajomstva je, že má byť vyriešené, po čom prestane byť tajomstvom. Vyvaruj sa tých, ktorí hovoria o otvorenej mysli a skromne priznávajú svoju nevedomosť. Je čas priznať svoju nevedomosť, a je čas svoju nevedomosť odložiť.

Druhou cnosťou je zriekanie sa. P. C. Hodgellová povedala: „To, čo môže byť zničené pravdou, nech je zničené.“[272] Necúvni pred skúsenosťami, ktoré môžu zničiť tvoje názory. Myšlienka, ktorú si nemôžeš pomyslieť, ťa ovláda viac než myšlienky, ktoré hovoríš nahlas. Podrob sa skúškam a prever sa v ohni. Zriekni sa emócie, ktorá spočíva na mylnom názore, a usiluj sa naplno precítiť tú emóciu, ktorá zodpovedá faktom. Ak sa k tvojej tvári blíži železo, o ktorom veríš, že je horúce, avšak ono je chladné, Cesta odporuje tvojmu strachu. Ak sa k tvojej tvári blíži železo, o ktorom veríš, že je chladné, avšak ono je horúce, Cesta odporuje tvojmu pokoju. Najprv vyhodnoť svoje názory a potom dospej k emóciám. Povedz si: „Ak je toto železo horúce, chcem veriť, že je horúce, a ak je chladné, chcem veriť, že je chladné.“ Vyvaruj sa pripútanosti k názorom, ktoré možno nechceš mať.

Treťou cnosťou je ľahkosť. Nech ťa vietor indícií unáša, akoby si bol lístkom bez vlastného smerovania. Vyvaruj sa zákopového boja proti indíciám, neochoty ustúpiť o každý meter, pocitu oklamania. Vzdaj sa pravde tak rýchlo, ako len dokážeš. Sprav to v tom okamihu, keď si uvedomíš, že vzdoruješ; v okamihu, keď vidíš, z ktorého kúta vietor indícií veje proti tebe. Buď neverný svojim stanoviskám a zraď ich silnejšiemu súperovi. Ak vnímaš indície ako obmedzenia, z ktorých sa snažíš oslobodiť, sám sa zväzuješ do reťazí vlastnej svojvôle. Nedokážeš totiž nakresliť správnu mapu mesta, ak sedíš vo svojej spálni so zatvorenými očami a impulzívne kreslíš čiary na papier. Musíš sa tým mestom prejsť a kresliť čiary zodpovedajúce tomu, čo vidíš. Ak si, vidiac mesto nejasne, myslíš, že môžeš posunúť čiaru čo len kúsok doprava, čo len kúsok doľava, podľa vlastného rozmaru, je to tá istá chyba.

Štvrtou cnosťou je vyrovnanosť. Ten, kto chce veriť, sa pýta: „Umožňujú mi tieto indície veriť?“ Ten, kto chce neveriť, sa pýta: „Nútia ma tieto indície veriť?“ Vyvaruj sa kladenia ťažkých bremien dôkazu iba na tie návrhy, ktoré sa ti nepáčia, a následnej obrany: „Je predsa dobré byť skeptický.“ Ak sa venuješ iba priaznivým indíciám, ktoré si vyberáš spomedzi zhromaždených údajov, potom čím viac údajov zhromažďíš, tým menej vieš. Ak si medzi argumentmi vyberáš, ktorých chyby budeš hľadať, alebo ako usilovne budeš tieto chyby hľadať, potom ťa každá chyba, ktorú sa naučíš odhaliť, urobí ešte hlúpejším. Ak si najprv naspodok papiera napíšeš „A preto je obloha zelená!“, nezáleží na tom, aké argumenty dodatočne napíšeš nad to; záver je už napísaný, a buď je už správny, alebo už nesprávny. Byť chytrým v argumentovaní nie je rozumnosť ale racionalizácia. Inteligencia, ak má byť na úžitok, musí byť použitá na niečo iné, ako na porazenie seba samej. Počúvaj hypotézy, ako pred teba predkladajú svoje prípady, ale pamätaj, že ty nie si hypotéza, ale sudca. Preto sa nepokúšaj argumentovať v prospech jednej alebo druhej strany, pretože keby si poznal svoj cieľ, už dávno by si tam bol.

→ http://lesswrong.com/lw/nc/newcombs_problem_and_regret_of_rationality/

272 Patricia C. Hodgell, *Seeker's Mask* (Meisha Merlin Publishing, Inc., 2001).

Piatou cnosťou je hádavosť. Tí, ktorí chcú zlyhať, musia najprv zabrániť svojim priateľov, aby im prišli na pomoc. Tí, ktorí sa múdro usmejú a povedia: „Ja sa nehádam“, odišli z dosahu pomoci a oddelili sa od spoločného úsilia. V hádke sa usiluj o presnú poctivosť, v záujme druhých i seba: Tá časť teba samého, ktorá prekrúca, čo hovoríš druhým, prekrúca aj tvoje vlastné myšlienky. Never tomu, že robíš druhým láskavosť, keď prijímaš ich argumenty; láskavosť robia oni tebe. Nemysli si, že férovosť voči všetkým stranám znamená postaviť sa presne doprostred medzi ich postoje; pravda nebýva na začiatku debaty rozdelená rovnakým dielom. Nemôžeš zvíťaziť vo faktickej otázke tým, že budeš bojovať päsťami alebo urážkami. Hľadaj test, ktorý umožní skutočnosti stať sa vaším sudcom.

Šiestou cnosťou je empirizmus. Korene poznania sú v pozorovaní, a jeho ovocím je predpoveď. Ktorý strom rastie bez koreňov? Ktorý strom nás živí bez ovocia? Ak strom padne v lese a nikto ho nepočuje, vydá pritom zvuk? Jeden povie: „Áno, vydá, pretože vzduch začne vibrovať.“ Iný povie: „Nie, nevydá, lebo nie je žiadne sluchové spracovanie v mozgu.“ Hoci sa hádajú, jeden hovorí „Áno“ a druhý hovorí „Nie“, obaja očakávajú v lese rovnakú skúsenosť. Nepýtaj sa, ktoré názory vyznávať, ale ktoré skúsenosti očakávať. Vždy vedz, o akom rozdieli v skúsenosti sa hádaš. Nedovoľ hádke, aby sa zatúlala a bola o niečom inom, napríklad o niekoho racionalistických cnostiach. Jerry Cleaver povedal: „To, čo ťa dostane, nie je chyba v aplikovaní nejakej náročnej, zamotanej, zložitej techniky. Je to prehliadnutie základných vecí. Nedržanie pohľadu na lopte.“[273] Nenechaj sa zaslepiť slovami. Keď odrátame slová, zostane očakávanie.

Siedmou cnosťou je jednoduchosť. Antoine de Saint-Exupéry povedal: „Dokonalosť sa dosahuje nie vtedy, keď nie je čo pridať, ale keď nie je čo odobrať.“[274] Jednoduchosť je cnosťou v názoroch, návrhoch, plánoch, a zdôvodneniach. Keď vyznávaš zložitý názor s mnohými malými podrobnosťami, každá ďalšia podrobnosť znamená ďalšiu šancu, že názor je nesprávny. Každé upresnenie pridáva k tvojmu bremenu; ak môžeš svoje bremeno uľahčiť, musíš tak urobiť. Každé steblo má moc zlomiť ti chrbát. O strojoch sa hovorí: Najspoľahlivejšia súčiastka je tá, ktorá v stroji vďaka jeho dizajnu nie je. O plánoch: Zamotaná sieť sa roztrhne. Reťaz z tisíca ohniviek ťa dovedie k správnemu záveru iba ak je každý krok správny; ak je však jeden krok nesprávny, môže ťa doviesť kamkoľvek. V matematika ani hora dobrých skutkov nedokáže odčiniť jeden hriech. Dávaj si preto pozor na každý krok.

Ôsmou cnosťou je pokora. Byť pokorný znamená prijať konkrétne opatrenia v očakávaní vlastných chýb. Priznať svoju omylnosť a potom s ňou nič neurobiť nie je pokorné; je to chvastanie sa svojou skromnosťou. Kto je najpokornejší? Tí, ktorí sa najšikovnejšie pripravujú na najhlbšie a najkatastrofálnejšie chyby vo svojich vlastných názoroch a plánoch. Keďže na tomto svete sú mnohí, ktorých od rozumnosti delí hlboká priepasť, začínajúci študenti rozumnosti vyhrávajú hádky a nadobúdajú prehnaný pohľad na svoje vlastné schopnosti. Byť lepším je však zbytočné: Život nie je známkovaný podľa stupnice. Najlepší fyzici starovekého Grécka nedokázali vypočítať dráhu padajúceho jablka. Neexistuje žiadna záruka, že tvoje najväčšie úsilie prinesie primeraný výsledok; preto neplytvaj myšlienkami na to, či sa iným darí horšie. Ak sa budeš porovnávať s ostatnými, neuvidíš omyly, ktoré majú všetci ľudia spoločné. Byť človekom znamená robiť desaťtisíc chýb. Nikto v tomto svete nedosahuje dokonalosť.

Deviatou cnosťou je perfekcionizmus. Čím viac chýb v sebe napravíš, tým viac ich ešte uvidíš. Čím tichšia bude tvoja myseľ, tým viac hluku budeš počuť. Keď si v sebe všimneš chybu, je to znamenie, že si pripravený hľadať posun na nasledujúcu úroveň. Ak budeš túto chybu tolerovať namiesto aby si ju napravil, neposunieš sa na nasledujúcu úroveň, ani nezískaš schopnosť všimnúť si nové chyby. V každom umení, ak nehľadáš dokonalosť, zastavíš sa ešte pred prvým krokom. Ak dokonalosť nie je možná, to ťa neospravedlňuje nepokúšať sa o ňu. Postav si najvyššiu latku, akú si vieš predstaviť, a potom hľadaj ešte vyššiu. Neuspokoj sa s odpoveďou, ktorá je takmer správna; hľadaj takú, ktorá je celkom správna.

Desiatou cnosťou je presnosť. Jeden príde a povie: Daný počet je medzi 1 a 100. Iný povie: Daný počet je medzi 40 a 50. Ak je daný počet 42, obaja majú pravdu, ale predpoveď toho druhého bola užitočnejšia a vystavila sa prísnejšej skúške. Čo je pravdou o jednom jablku, nemusí byť pravdou o

273 Cleaver, *Immediate Fiction: A Complete Writing Course.*

274 Antoine de Saint-Exupery, *Terre des Hommes* (Paris: Gallimard, 1939).

druhom jablku; preto možno o jednom jablku povedať viac než o všetkých jablkách na celom svete. Najužšie výroky režú najhlbšie; nôž krája ostrou hranou. Čo platí pre mapu, platí aj pre umenie kreslenia máp: Cesta je presným Umením. K pravde sa nekráča, ale tancuje. V každom jednom kroku tanca tvoja noha presne došliapne na správne miesto. Každý kus indície posúva tvoje názory presne o správne množstvo, ani viac, ani menej. Aké je správne množstvo? Aby si to spočítal, musíš študovať teóriu pravdepodobnosti. Aj keď nevieš počítať matematické rovnice, vedomosť, že existujú, ti hovorí, že krok tanca je presný a niet v ňom miesta pre tvoje rozmary.

Jedenástou cnosťou je učenosť. Študuj mnohé vedy a nasaj ich silu ako svoju vlastnú. Každá oblasť, ktorú stráviš, ťa spraví väčším. Ak pohltíš dostatok vied, medzery medzi nimi sa zmenšia a tvoje poznanie sa stane súvislým celkom. Ak budeš nenásytný, môžeš byť väčším než hory. Osobitne dôležité je konzumovať matematiku a vedy súvisiace s rozumnosťou: Evolučnú psychológiu, heuristiku a skreslenia, sociálnu psychológiu, teóriu pravdepodobnosti, teóriu rozhodovania. Nesmú to však byť jediné veci, ktoré študuješ. Umenie musí mať svoj cieľ mimo seba, inak sa zrúti do nekonečnej rekurzie.

Pred týmito jedenástimi cnosťami je cnosť, ktorá je bez mena.

Miyamoto Musashi napísal v *Knihe piatich kruhov*:[275]

> „Prvoradou vecou, keď uchopíš meč do rúk, je tvoj zámer ťať nepriateľa, akýmkoľvek spôsobom. Kedykoľvek odrážaš, udieraš, bodáš, sekáš, alebo sa dotýkaš nepriateľovho tnúceho meča, musíš v tom istom pohybe ťať nepriateľa. Je podstatné toto dosiahnuť. Ak myslíš iba na odrážanie, udieranie, bodanie, sekanie, alebo dotýkanie sa nepriateľa, nedokážeš ho naozaj sťať. Viac než čokoľvek iné, musíš myslieť na to, ako svoj pohyb premeníš na jeho tnutie.“

Každý krok tvojho uvažovania musí v tom istom pohybe ťať do správnej odpovede. Viac než na čokoľvek iné musíš myslieť na to, ako premeniť svoju mapu, aby odrážala územie.

Ak sa ti nepodarí dosiahnuť správnu odpoveď, márne je namietať, že si konal primerane.

Ako môžeš zlepšiť svoje chápanie rozumnosti? Nie tak, že si budeš hovoriť: „Je mojou povinnosťou byť rozumný.“ Tým by si iba uchovával svoje mylné chápanie. Možnože chápeš rozumnosť tak, že je rozumné veriť slovám Veľkého Učiteľa, a Veľký Učiteľ hovorí: „Obloha je zelená,“ a ty sa pozrieš na oblohu a vidíš modrú. Ak si pomyslíš: „Môže sa zdať, že obloha je modrá, ale rozumné je veriť slovám Veľkého Učiteľa,“ stratíš šancu na objavenie svojej chyby.

Nepýtaj sa, či je „správna Cesta“ robiť toto alebo tamto. Pýtaj sa, či je obloha modrá alebo zelená. Ak budeš priveľa hovoriť o Ceste, nedosiahneš ju.

Môžeš skúšať nazvať tento najvyšší princíp menami ako „mapa, ktorá odráža územie“ alebo „skúsenosť úspechu a zlyhania“ alebo „Bayesovská teória rozhodovania“. Ale možno si túto cnosť bez mena popísal nesprávne. Ako odhalíš svoju chybu? Nie porovnávaním tohto popisu so sebou samým, ale porovnaním ho s tým, ktoré si nepomenoval.

Ak budeš veľa rokov precvičovať techniky a podrobíš sa prísnym obmedzeniam, možno raz zazrieš stred. Potom uvidíš, ako všetky tieto techniky sú jednou technikou, a budeš sa pohybovať správne bez pocitu obmedzenia. Musashi napísal: „Keď oceníte moc prírody a spoznáte rytmus každej situácie, budete vedieť nepriateľa prirodzene udrieť a prirodzene seknúť. Toto všetko je Cesta Prázdnoty.“

Toto je teda dvanásť cností rozumnosti:

Zvedavosť, zriekanie sa, ľahkosť, vyrovnanosť, hádavosť, empirizmus, jednoduchosť, pokora, perfekcionizmus, presnosť, učenosť, a prázdnota.

* →

275 Musashi, *Book of Five Rings*.

→ http://www.yudkowsky.net/rational/virtues/

Kniha VI.

Stať sa silnejším

Začiatky: Úvod

(napísal Rob Bensinger)

Táto posledná kniha z celku *Rozumnosť: od algoritmov po zombie* nie je ani tak záverom, ako výzvou konať. V duchu *stávania sa silnejším* ako odrazový most k ďalšiemu skúmaniu zakončím citovaním zdrojov, ktoré čitateľ môže použiť, aby sa dostal za tieto postupnosti a našiel plnšie pochopenie bayesiánstva.

V tomto texte sa použitá definícia normatívnej rozumnosti podľa bayesovskej teórie pravdepodobnosti a teória rozhodovania je štandardom v kognitívnej vede. Ako úvod do heuristík a skreslení si pozrite *Myslenie a rozhodovanie* od Barona.[276] Ako všeobecný úvod do oblasti si pozrite *Oxfordskú príručku o myslení a usudzovaní*.[277]

Argumenty o *filozofii* rozumnosti, ktoré nájdete na týchto stránkach, sú kontroverznejšie. Yudkowsky napríklad tvrdí, že rozumný agent by si mal v Newcombovom probléme vybrať jednu krabicu – čo je menšinový názor medzi teoretikmi rozhodovania.[278] (Netechnický popis Newcombovho problému uvádza Holt.[279]) *Dobré a skutočné* od Garyho Dreshera nezávisle dochádza k mnohým rovnakým záverom ako Yudkowsky o filozofii vedy a teórii rozhodovania.[280] Táto kniha teda poslúži ako výborné pojednanie základného filozofického obsahu celku *Rozumnosť: od algoritmov po zombie*.

Talbott rozlišuje niekoľko pohľadov v rámci bayesovskej epistemológie, vrátane postoja E. T. Jaynesa, že nie všetky východiskové predpoklady sú rovnako rozumné.[281,282] Podobne ako Jaynes, aj Yudkowsky sa zaujíma o doplnenie bayesovského kritéria optimálnej zmeny názoru kritériom optimálneho východiska. Toto Yudkowskeho spája s výskumníkmi snažiacimi sa lepšie pochopiť všeobecnú UI pomocou vylepšenej teórie ideálneho usudzovania, ako napríklad Marcus Hutter.[283] Dlhšiu diskusiu o filozofických snahách naturalizovať teórie poznania uvádza Feldman.[284]

„Bayesiánstvo“ sa často stavia do kontrastu s „frekventizmom“. Niektoré frekventisti kritizujú bayesovcov za zaobchádzanie s pravdepodobnosťami ako so subjektívnymi stavmi názoru, namiesto ako s objektívnymi frekvenciami udalostí. Kruschke a Yudkowsky na to odpovedajú, že frekventizmus je ešte „subjektívnejší“ než bayesiánstvo, pretože frekventistické priradenia pravdepodobnosti závisia od zámerov experimentátora.[285]

Dôležité je, že tieto filozofické nezhody by sme si nemali mýliť s rozdielom medzi bayesovskými a frekventistickými metódami analýzy údajov, ktoré môžu byť obe užitočné, ak ich použijeme správne. Používanie bayesovských štatistických nástrojov sa od 80. rokov stalo lacnejšie, ich informatívnosť, intuitívnosť a všeobecnosť si získali širšie uznanie, čoho výsledkom bola „bayesiánska revolúcia“ v mnohých vedách. Tradičné frekventistické metódy sú však naďalej obľúbenejšie, a v niektorých

276 Jonathan Baron, *Thinking and Deciding* (Cambridge University Press, 2007).

277 Keith J. Holyoak and Robert G. Morrison, *The Oxford Handbook of Thinking and Reasoning* (Oxford University Press, 2013).

278 Bourget and Chalmers, „What Do Philosophers Believe?“

279 Holt, „Thinking Inside the Boxes.“

280 Gary L. Drescher, *Good and Real: Demystifying Paradoxes from Physics to Ethics* (Cambridge, MA: MIT Press, 2006).

281 William Talbott, „Bayesian Epistemology,“ in *The Stanford Encyclopedia of Philosophy*, Fall 2013, ed. Edward N. Zalta.

282 Jaynes, *Probability Theory*.

283 Marcus Hutter, *Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability* (Berlin: Springer, 2005), doi:10.1007/b138233.

284 Richard Feldman, „Naturalized Epistemology,“ in *The Stanford Encyclopedia of Philosophy*, Summer 2012, ed. Edward N. Zalta.

285 John K. Kruschke, „What to Believe: Bayesian Methods for Data Analysis,“ *Trends in Cognitive Sciences* 14, no. 7 (2010): 293–300.

kontextoch sú stále jasne lepšie než bayesiánske prístupy. Zábavný a čitateľný úvod do tejto témy ponúka *Robenie bayesovskej analýzy údajov* od Kruschkeho.[286]

Vo svetle indícií, že výcvik v štatistike – a niektorých ďalších oblastiach, napríklad v psychológii – zlepšuje usudzovacie schopnosti mimo učebne, štatistická gramotnosť priamo súvisí s projektom prekonávania skreslení. (Lekcie formálnej logiky a neformálnych omylov sa neukázali ako podobne úspešné.)[287,288]

Umenie vo svojich začiatkoch

Uzatvárame troma postupnosťami o individuálnom a kolektívnom sebazdokonaľovaní. „Yudkowskeho dospievanie“ poskytuje poslednú hĺbkovú ilustráciu dynamiky nerozumného názoru, tentokrát so zameraním na autorovu vlastnú intelektuálnu históriu. „Pokúsiť sa o zložité“ sa pýta, čo treba, aby sme vyriešili naozaj zložitý problém – vrátane požiadaviek, ktoré idú za epistemickú rozumnosť. Nakoniec, „Remeslo a komunita“ pojednáva o racionalistických skupinách a skupinovej rozumnosti, kladúc si otázky:

- Možno sa rozumnosť naučiť, a učiť ju?
- Ak áno, aké zdokonalenie je možné?
- Ako si môžeme byť istí, že vidíme skutočný výsledok vyučovania rozumnosti, a že si vyberáme správny cieľ?
- Aké spoločenské normy by tento proces sebazdokonaľovania zjednodušili?
- Dokážeme účinne spolupracovať na veľkých problémoch bez obetovania svojej sloboby myslenia a konania?

A najmä: Čo tu chýba? Čo by sa malo nachádzať v ďalšej generácii úvodov do rozumnosti – v tých, ktoré nahradia tento text, vylepšia jeho štýl, otestujú jeho odporúčania, doplnia jeho obsah, a vyrastú celkom novými smermi?

Hoci Yudkowskeho pohli napísať tieto eseje jeho vlastné filozofické chyby a profesionálne ťažkosti v teórii UI, výsledný materiál sa ukázal byť užitočný pre omnoho širšie publikum. Pôvodné články na blogu inšpirovali rast stránky *Less Wrong*, spoločenstva intelektuálov a kutilov života so spoločným záujmom o kognitívnu vedu, informatiku a filozofiu. Yudkowsky a ďalší autori na *Less Wrong* pomohli založiť hnutie efektívneho altruizmu, živé a odvážne úsilie identifikovať humanitárne charity a kauzy s najväčším dopadom. Tieto texty podnietili aj založenie Centra pre aplikovanú rozumnosť, neziskovej organizácie, ktorá sa snaží preložiť výsledky z vedy o rozumnosti do užitočných techník na sebazdokonaľovanie.

Neviem, čo bude ďalej – aké ďalšie nekonvenčné projekty alebo myšlienky by mohli čerpať inšpiráciu z týchto stránok. Iste nám nechýba dostatok globálnych problémov, a umenie rozumnosti je nová a nedokončená vec. Nie je veľa racionalistov, a mnoho vecí je nedorobených.

Kamkoľvek však zamieriš ďalej Ty, čitateľ – kiež dobre poslúži Tvojim cieľom.

---

286 John K. Kruschke, *Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan* (Academic Press, 2014).

287 Geoffrey T. Fong, David H. Krantz, and Richard E. Nisbett, „The Effects of Statistical Training on Thinking about Everyday Problems,“ *Cognitive Psychology* 18, no. 3 (1986): 253–292, doi:10.1016/0010-0285(86)90001-0.

288 Paul J. H. Schoemaker, „The Role of Statistical Knowledge in Gambling Decisions: Moment vs. Risk Dimension Approaches,“ *Organizational Behavior and Human Performance* 24, no. 1 (1979): 1–17.

X: Yudkowskeho dospievanie

292. Špirála smrti môjho detstva

Moji rodičia mali vždy sklon bagatelizovať hodnotu inteligencie. A vyzdvihovať hodnotu... úsilia, ako odporúčajú najnovšie výskumy? Nie, nie úsilia. *Skúsenosti.* Pekné nedosiahnuteľné kladivo, ktorým môžete zraziť hrebienok nadanému mladému dieťaťu, iste. To bolo to, čo mi moji rodičia povedali, keď som napríklad pochyboval o židovskom náboženstve. Pokúšal som sa zostaviť argument, a bolo mi povedané niečo v duchu: „Logika má svoje hranice, keď budeš starší, pochopíš, že skúsenosť je dôležitá, a potom uvidíš pravdu judaizmu.“ Viac som to neskúšal. Jedenkrát som skúsil pochybovať o judaizme v škole, bol som usadený, druhýkrát som to už neskúšal. Nikdy som nebol nechápavým žiakom.

Kedykoľvek moji rodičia robili niečo nerozumné, vždy to bolo: „My vieme lepšie, pretože máme viac skúseností. Pochopíš to, keď budeš starší: zrelosť a múdrosť sú dôležitejšie než inteligencia.“

Ak bola toto snaha upriamiť mladého Eliezera na *inteligenciu uber alles*, bol to ten absolútne najúspešnejší príklad reverznej psychológie, o akom som kedy počul.

Lenže moji rodičia neboli takíto prefíkaní, a výsledky neboli celkom pozitívne.

Dlhú dobu som si myslel, že ponaučením z tohto príbehu je, že skúsenosť sa nemôže porovnávať s čírou živelnou inteligenciou. Bolo to až omnoho neskôr, po dvadsiatke, keď som sa obzrel a uvedomil som si, že som *pred pubertou* nemohol byť inteligentnejší než moji rodičia, keď môj mozog ešte nebol celkom vyvinutý. Vo veku jedenásť rokov, keď už som bol takmer úplný ateista, by som nemohol poraziť svojich rodičov v žiadnom *férovom* zápase myslí. Moje SAT skóre bolo na 11-ročného vysoké, ale nebolo by vyššie než SAT skóre mojich rodičov v dospelosti. Vo férovom súboji by inteligencia a skúsenosti mojich rodičov úplne prevalcovali ľubovoľné dieťa v puberte. Bola to dysracionalita, čo ich porazilo; používali svoju inteligenciu iba na to, aby porazili sami seba.

Lenže *toto* porozumenie prišlo omnoho neskôr, keď moja inteligencia spracovala a destilovala omnoho viac rokov skúsenosti.

Ponaučenie, ktoré som si odvodil, keď som bol mladý, bolo že každý, kto bagatelizuje hodnotu inteligencie, vôbec nerozumie inteligencii. Moja vlastná inteligencia ovplyvnila každú stránku môjho života a mysle a osobnosti; to bolo masívne zrejmé, keď som sa obzrel. „Inteligencia nemá nič spoločné s múdrosťou alebo s tým, či je niekto dobrý človek“ - ale, vari poznanie seba samého nemá nič spoločné s múdrosťou alebo s tým, či je niekto dobrý človek? Modelovať sám seba vyžaduje inteligenciu. Napríklad, na štúdium evolučnej psychológie potrebujete dosť inteligencie.

My *sme* také karty, aké nám boli rozdané, a inteligencia je tá najneférovejšia zo všetkých týchto kariet. Omnoho neférovejšia než bohatstvo alebo zdravie alebo krajina pôvodu, neférovejšia než prirodzená hladina šťastia. Ľudia majú problém zmieriť sa s tým, že život môže byť neférový, nie je to štastná myšlienka. „Inteligencia nie je taká dôležitá ako X“ je jeden spôsob, ako sa vyhnúť tejto neférovosti, odmietnuť sa ňou zaoberať, myslieť namiesto toho na niečo veselšie. Je to pokušenie, aj pre tých, ktorí dostali zlé karty, aj pre tých, ktorí dostali dobré. Rovnako ako bagatelizovať dôležitosť peňazí je pokušením aj pre chudobných aj pre bohatých.

Lenže mladý Eliezer bol transhumanista. Rozdať svoje IQ bude vyžadovať omnoho viac práce než keby som sa bol narodil s hromadou peňazí. Ale je to riešiteľný problém, ktorému treba čeliť priamo a vyriešiť ho. Aj keby to zabralo celý môj život. „Silní existujú, aby slúžili slabým,“ napísal mladý Eliezer, „a tejto povinnosti sa môžu zbaviť iba tak, že urobia druhých rovnako silnými.“ Liezli mi na nervy Randovské a Nietzscheovské trendy v sci-fi, a ako ste už asi pochopili, mladý Eliezer mal sklon brať veci *príliš ďaleko do opačného extrému*. Nikto neexistuje iba preto, aby slúžil. Ale ja som to skúsil a neľutujem. Ak toto nazvete puberťáckym bláznovstvom, je zriedkavé vidieť, ako sa dospelej múdrosti darí lepšie.

Každý potrebuje viac inteligencie. Vrátane mňa, zdôrazňoval som starostlivo. Bolo mi vzdialené vyhlasovať nový svetový poriadok so sebou navrchu – to bolo to, čo by urobil stereotypný darebák zo

science fiction, alebo horšie, typický puberťák, a ja som si nikdy nedovolil takéto klišé. Nie, *každý* potrebuje byť múdrejší. Sme všetci na jednej lodi: Pekná, povznášajúca myšlienka.

Eliezer$_{1995}$ prečítal veľa science fiction. Mal morálku, etiku, a dokázal vidieť tie samozrejmejšie pasce. Žiadne zoznamy *Homo novis*. Žiadna čiara medzi ním a ostatnými. Žiadna sofistikovaná filozofia, ktorý by ho postavila na vrch hromady. To bolo príliš samozrejmé zlyhanie. Áno, dával si pozor na to, aby aj sám seba nazýval hlúpym, a nikdy netvrdil, že je morálne nadradený. Nuž, ani teraz to nevidím inak, hoci už zo svojej etiky nerobím toľkú dramatickú produkciu. (Alebo možno by bolo presnejšie povedať, že som prísnejší na to, kedy si na chvíľku dovolím sám sebe zablahoželať.)

Toto všetko hovorím, aby som zdôraznil, že Eliezer$_{1995}$ nebol taký nedôstojný, aby zlyhal nejakým *samozrejmým* spôsobom.

A potom Eliezer$_{1996}$ natrafil na pojem Singularity. Bol to blesk zjavenia? Vyskočil som z kresla a kričal: „Eurisko!“? Nie, nie. Nebol som taký divadelník. Akurát mi prišlo ohromne samozrejmé v spätnom pohľade, že inteligencia väčšia než ľudská zmení budúcnosť podstatnejšie než hocijaká hmotná veda. A hneď som vedel, že *toto* budem robiť po celý zvyšok svojho života, vytvárať Singularitu. Nie nanotechnológiu, ako som si myslel, keď som mal jedenásť rokov; nanotechnológia by bola iba nástroj vytvorený inteligenciou. Inteligencia bola predsa ešte *mocnejšia*, ešte väčšie požehnanie, než som si uvedomoval predtým.

Bola toto veselá špirála smrti? Ako sa neskôr ukázalo, áno: viedlo to k prijatiu aj *falošných* veselých myšlienok o inteligencii. Možno by ste mohli nakresliť čiaru v bode, keď som začal veriť, že pre superinteligenciu určite nebude prekážkou ani hranica rýchlosti svetla.

(Ako sa moje názory na inteligenciu zmenili odvtedy... pozrime sa: Keď dnes pomyslím na zlé karty rozdané ľuďom, ako prvé pomyslím na smrť a na starobu. Každý musí mať nejakú úroveň inteligencie, takú alebo onakú, a dôležitá vec z hľadiska teórie zábavy je, že by časom mala *rásť*, a nie klesať ako teraz. Nie je toto chytrý spôsob, ako sa cítiť lepšie? Lenže ja sa dnes toľko nesnažím bagatelizovať svoju vlastnú inteligenciu, pretože to je len iný spôsob, ako na ňu upútavať pozornosť. Na ľudské pomery som bystrý, keby prišlo na túto tému, a aký mám z toho pocit, to je moja vec.

Tá časť, že inteligencia je páka, ktorá dvíha svety, je stále rovnaká. Až na to, že inteligencia mi začala byť menej tajomnou, takže teraz jasnejšie vidím inteligenciu ako niečo zhmotnené vnútri fyziky. Superinteligencia možno bude cestovať rýchlejšie ako svetlo, ak sa ukáže, že to skutočné fyzikálne zákony umožňujú, ale ak nie, tak potom nebude. Nie je to nemysliteľné, ale nestavil by som na to.)

Ale tá skutočne zlá odbočka prišla neskôr, v bode, keď niekto povedal: „Hej, odkiaľ vieš, že tá superinteligencia bude morálna? Inteligencia nemá nič spoločné s tým, či si dobrý človek, vieš – tomu sa hovorí múdrosť, mladý génius.“

A hľa, mladému Eliezerovi sa zdalo samozrejmé, že toto je púhe popieranie. Jeho vlastný starostlivo zostavený etický kódex bol iste zložený pomocou jeho inteligencie, a spočíval na základoch jeho inteligencie. Hocijaký blázon môže vidieť, že inteligencia veľmi súvisí s etikou, morálkou a múdrosťou; však skúste vysvetliť Väzenskú dilemu šimpanzovi, nie?

Iste teda by superinteligencia nevyhnutne znamenala supermorálku.

Preto sa hovorí: „Rodičia robili všetky tie veci, ktoré hovoria svojim deťom, aby ich nerobili, veď odtiaľ vedia, že sa to robiť nemá.“

*→

293. Moja najlepšia a najhoršia chyba

V predchádzajúcej kapitole som opísal afektívnu špirálu smrti mladého Eliezera ohľadom niečoho, čo nazýval „inteligencia“. Eliezer$_{1996}$, dokonca aj Eliezer$_{1999}$ by odmietol skúsiť napísať matematickú definíciu – vedome, úmyselne odmietol. Nebol by ochotný priradiť „inteligencii“ vôbec nejakú definíciu.

→ http://lesswrong.com/lw/ty/my_childhood_death_spiral/

Prečo? Pretože existuje je klasický chyták v UI, kde si definujete „inteligenciu“ aby znamenala niečo ako „logické uvažovanie“ alebo „schopnosť vzdať sa predchádzajúcich záverov, keď už viac neplatia“ a potom postavíte nejaký lacný dokazovač tvrdení alebo ad-hoc nemonotónny rozhodovač, a poviete: „Hľa, naprogramoval som inteligenciu!“ Ľudia prichádzajú so zlými definíciami inteligencie – sústredia sa na koreláty a nie na jadro – a potom naháňajú tie povrchné definície, ktoré si napísali, a zabúdajú na, veď viete, skutočnú *inteligenciu*. Eliezer$_{1996}$ si nehodlal postaviť kariéru v oblasti umelej inteligencie. Chcel iba myseľ, ktorá by naozaj dokázala postaviť nanotechnológiu. Nebol teda v pokušení predefinovať inteligenciu, aby mohol uverejniť odborný článok.

Keď sa obzriem, zdá sa mi, že mnohé z mojich chýb možno definovať ako zájdenie príliš ďaleko opačným smerom pri pohľade na hlúposť niekoho iného: Keď som tak často videl zneužívanie pokusov o definíciu „inteligencie“, odmietol som ju vôbec definovať. Čo keby som povedal, že inteligencia je X, a *v skutočnosti* by to nebolo X? Vedel som v intuitívnom zmysle, čo hľadám – niečo dosť mocné na to, aby rozobralo hviezdy kvôli surovinám – a nechcel som padnúť do pasce, že sa od tohto nechám odlákať definíciami.

Podobne som videl, ako tak veľa projektov UI stroskotalo na závisti voči fyzike – snažili sa držať jednoduchej a elegantnej matematiky, a v dôsledku toho sa obmedzovali na hračkárske systémy – zovšeobecnil som si, že žiadna matematika dostatočne jednoduchá na to, aby sa dala formalizovať peknou rovnicou, pravdepodobne nebude fungovať pre, veď viete, *skutočnú* inteligenciu. „Okrem Bayesovej vety“, dodal Eliezer$_{2000}$; čo podľa toho, ako sa pozeráte, buď zmierňuje celkovú závažnosť jeho priestupku, alebo ukazuje, že mal začať podozrievať celé zovšeobecnenie namiesto snahy pridať jedinú výnimku.

Ak sa čudujete, prečo si Eliezer$_{2000}$ niečo takéto myslel – neveril v matematiku inteligencie – no, je pre mňa ťažké si na to spomenúť. Iste to nebolo preto, že by som nemal rád matematiku. A keby som mal ukázať na základnú príčinu, bolo by to čítanie príliš málo, príliš populárnych, a nesprávnych kníh o umelej inteligencii.

Lenže vtedy som si nemyslel, že odpovede prídu z oblasti umelej inteligencie; prevažne som ju odpísal ako choré, mŕtve odvetvie. Nie je teda prekvapujúce, že som strávil tak málo času jej skúmaním. Veril som v klišé, že umelá inteligencia dáva prehnané sľuby. To sa dá zaradiť do vzorca „príliš ďaleko opačným smerom“ - odvetvie nesplnilo svoje sľuby, preto som ho bol pripravený odpísať. V dôsledku toho som nehľadal dosť usilovne na to, aby som našiel matematiku, ktorá by nebola falošná.

Moja mladícka nedôvera v matematiku všeobecnej inteligencie bola súčasne jednou z mojich celkovo najhorších chýb, a jednou z mojich celkovo najlepších chýb.

Pretože som neveril, že môže existovať nejaká jednoduchá odpoveď na inteligenciu, šiel som čítať o kognitívnej psychológii, funkčnej neuroanatómii, výpočtovej neuroanatómii, evolučnej psychológii, evolučnej biológii, a viac než jednej vetve umelej inteligencie. Keď som mal niečo, čo vyzeralo ako jednoduchý jasný nápad, nezastavil som sa tam, ani som sa nerozbehol, aby som to implementoval, pretože som vedel, že aj keby to bola pravda, aj keby to bolo nevyhnutné, nebude to dostatočné: inteligencia nemala byť niečo jednoduché, nemala mať odpoveď, ktorá sa zmestí na tričko. Mala to byť veľká skladačka s mnohými kúskami; a keď nájdete jeden kúsok, nerozbehnete sa držiac ho triumfálne vysoko, ale hľadáte ďalej. Pokúste sa postaviť myseľ, v ktorej chýba jediný kúsok, a možno sa nestane nič zaujímavé.

Mýlil som sa v názore, že akademická oblasť umelej inteligencie je mŕtva pustatina; a ešte viac v názore, že nemôže existovať matematika inteligencie. Ale neľutujem, že som študoval napríklad funkčnú neuroanatómiu, dokonca aj keď si *dnes* myslím, že umelá inteligencia by vôbec nemala vyzerať ako ľudský mozog. Štúdium neuroanatómie znamenalo, že som získal predstavu, že ak rozdelíte myseľ na časti, tie časti budú niečo ako „zraková kôra“ a „mozoček“ - a nie „modul na obchodovanie na burze“ alebo „modul uvažovania zdravým rozumom“, čo je štandardná nesprávna cesta v UI.

Štúdium oblastí ako funkčná neuroanatómia a kognitívna psychológia mi dalo veľmi odlišnú predstavu o tom, ako musia vyzerať mysle, než by ste dostali z čítanie kníh o UI – dokonca aj dobrých kníh o UI.

Keď vymažete všetky tie nesprávne závery a nesprávne zdôvodnenia, a opýtate sa iba, čo tento názor spôsobil, že mladý Eliezer naozaj *urobil*...

Potom názor, že umelá inteligencia je chorá a že skutočná odpoveď musí prísť zo zdravších odvetví mimo, ho viedol študovať veľa kognitívnej vedy;

Názor, že UI nemôže mať jednoduché odpovede, ho viedol nezastaviť sa predčasne na jednej skvelej myšlienke, ale zhromaždiť veľa informácií;

Názor, že je lepšie nedefinovať inteligenciu, viedol k situácii, kde študoval daný problém roky predtým, než začal navrhovať systematizáciu.

O tomto hovorím, keď poviem, že je to jedna z mojich celkovo najlepších chýb.

Keď sa po rokoch obzriem naspäť, odvodil som si v tomto zmysle veľmi silné ponaučenie:

Čo nakoniec naozaj urobíte, zatieňuje chytré dôvody, prečo ste to robili.

Porovnajte si úžasne chytré uvažovanie, ktoré vás vedie k štúdiu mnohých vied, s úžasne chytrým uvažovaním, ktoré hovorí, že všetky tie knihy netreba čítať. Dodatočne, keď sa ukáže, že vaše úžasne chytré uvažovanie bolo vlastne hlúpe, skončíte v omnoho lepšej situácii, ak to bolo úžasne chytré uvažovanie toho prvého typu.

Keď sa obzriem na svoju minulosť, zaskočí ma množstvo polo-náhodných úspechov, množstvo prípadov, keď som urobil niečo správne z nesprávneho dôvodu. Z vášho pohľadu by ste toto mohli pripísať antropickému princípu: keby som uviazol v skutočnej slepej uličke, pravdepodobne by ste nečítali moje veci na tomto blogu. Z môjho pohľadu tu zostávajú akési rozpaky. Moja Tradičná Racionalistická výchova mi poskytla veľa riadených sklonov k týmto „náhodným úspechom“ - vychýlila ma smerom k racionalizovaniu dôvodov, prečo študovať, namiesto prečo neštudovať, zabránila mi úplne sa stratiť, pomohla mi zotaviť sa z chýb. Napriek tomu nič z toho nebolo správne konanie zo správneho dôvodu, a to je strašidelná vec, keď sa obzriete a uvidíte to ako históriu svojej mladosti. Jedným z mojich hlavných dôvodov písania na *Overcoming Bias* je zanechať trasu tam, kde som sa ocitol náhodou – aby som sa urobil zbytočnou rolu, ktorú zohralo šťastie v mojom vlastnom staní sa racionalistom.

Takže čo to robí jednou z mojich celkovo najhorších chýb? Pretože niekedy „neformálne“ je len iný spôsob ako povedať „podľa nízkej latky kvality“. Mal som úžasne chytré dôvody, prečo je okej, keď nedefinujem presne „inteligenciu“ a takisto niekoľko mojich ďalších pojmov: konkrétne, že iní ľudia pri snahe definovať ich zablúdili. To bola brána, cez ktorú mohlo vstúpiť lajdácke uvažovanie.

Mal som teda vyskočiť a pokúsiť sa hneď zostaviť správnu definíciu? Nie, všetky dôvody, prečo som vedel, že toto je nesprávne, platili; nemôžete si vycucať správnu definíciu z prsta, keď nemáte primerané vzdelanie.

Nemôžete dostať *správnu* definíciu ohňa, ak neviete o atómoch a molekulách; budete na tom lepšie, keď poviete: „táto oranžová jasná vec“. A musíte byť schopní o tejto oranžovej jasnej veci hovoriť, aj keď nedokážete presne povedať, čo to je, ak chcete oheň skúmať. Ale dnes by som povedal, že všetko uvažovanie na tejto úrovni je niečo, čomu nemôžete dôverovať – je to skôr niečo, čo robíte na ceste k lepšiemu poznaniu, ale *nedôverujete* tomu, *neukladáte na to svoju váhu*, neodvodzujete z toho pevné závery, bez ohľadu na to, aké nevyhnutné vyzerá toto neformálne uvažovanie.

Mladý Eliezer položil svoju váhu na nesprávnu dlaždicu na podlahe – stúpil do nastraženej pasce.

* →

294. Vychovaný v technofílii

Môj otec zvykol hovoriť, že keby dnešný systém existoval pred sto rokmi, automobily by boli postavené mimo zákon, aby sme ochránili náš konský priemysel.

→ http://lesswrong.com/lw/tz/my_best_and_worst_mistake/

Jedným z hlavných vplyvov môjho detstva bolo čítanie *O krok ďalej von* od Jerryho Pournellea vo veku deväť rokov. Bola to Pournelleova reakcia na Paula Ehrlicha a Rímsky klub, ktorí v 1960-tych a 1970-tych rokoch hovorili, že sa na Zemi míňajú suroviny a masívne hladomory sú vzdialené iba pár rokov. Bola to reakcia na takzvaný štvrtý zákon termidynamiky Jeremyho Rifkina; bola to reakcia na všetkých ľudí, ktorí sa báli jadrovej energie a snažili sa ju regulovať až do zániku.

Vyrástol som vo svete, kde hranice medzi Dobrými Chlapmi a Zlými Chlapmi boli nakreslené celkom jasne; nie apokalyptická záverečná bitka, ale bitka, ktorú treba vybojovať znovu a znovu, bitka, ktorej historické ozveny sa vracajú späť až po priemyselnú revolúciu, a kde môžete zhromaždiť historické indície o skutočných výsledkoch.

Na jednej strane boli vedci a inžinieri, ktorí priniesli všetko zvyšovanie životného štandardu od temných čias, ktorých práca umožnila luxusné veci ako je demokracia, vzdelané obyvateľstvo, stredná trieda, zrušenie otroctva.

Na druhej strane tí, ktorí kedysi protestovali voči očkovaniu proti kiahňam, anestetikám počas pôrodu, parným strojom, a heliocentrizmu: Teológia, ktorí vzývali k návratu do dokonalej doby, ktorá nikdy neexistovala, starí bieli muži v politike držiaci sa svojich zabehaných spôsobov, skupiny špeciálnych záujmov ktoré mali čo stratiť, a mnohí, pre ktorých bola veda zavretou knihou a báli sa toho, čomu nerozumeli.

A tí, čo predstierali Hlbokú Múdrosť, sa snažili hrať na strednú cestu, prednášali uložené myšlienky o tom, že technológia pomáha ľudstvu alebo iba ak je primerane regulovaná – tvrdili v rozpore s holými historickými faktmi, že samotná veda nie je ani dobrá ani zlá – zostavovali slávnostne vyzerajúce úradné komisie, aby dali ostentatívne najavo svoju opatrnosť – a potom čakali na potlesk. Akoby pravda bola vždy kompromisom. Akoby niekto mohol naozaj vidieť tak ďaleko dopredu. Bolo by ľudstvo na tom lepšie, keby sme mali úprimne znepokojenú verejnú diskusiu o používaní ohňa, a keby na jeho využívanie dozerali zostavené komisie?

Keď som vstúpil do tohto problému, začal som mať alergiu na všetko, čo zodpovedalo vzorcu: „Ach, ale technológia má aj svoje riziká, nielen výhody, môj maličký.“ Uplatňoval som prezumpciu viny, že sa buď snažíte zožať nejaký lacný potlesk, alebo sa skryto pokúšate regulovať technológiu až k jej zániku. A v oboch prípadoch ignorujete historický záznam, ktorý je výrazne *v prospech* technológií, ktorých sa ľudia kedysi báli.

Dnes Robin Hanson nadhodil tému, ako FDA pomaly schvaľuje lieky schválené v iných krajinách. Niekto v komentároch podotkol, že Thalidomid sa predával v 50 krajinách pod 40 názvami, ale iba malé množstvo sa dostalo do USA, takže sa na celom svete narodilo 10 000 poškodených detí, ale iba 17 detí v USA.

Lenže koľko ľudí zomrelo kvôli pomalému schvaľovaniu v USA, keď potrebné lieky rýchlejšie schválili v iných krajinách – všetky tie lieky, ktoré *nedopadli* zle? A túto otázku sa pýtam, pretože je to niečo, o čom môžeme skúsiť pozbierať štatistiky – ale nepovie to nič o všetkých tých liekoch, ktoré ani nikdy neboli *vyvinuté*, pretože proces schvaľovania je taký dlhý a drahý. Podľa tohto zdroja dlhší proces schvaľovania FDA zabráni ročne 5 000 úmrtiam, keď zastaví lieky, ktoré sa neskôr ukážu ako škodlivé, ale spôsobí aspoň 20 000 – 120 000 úmrtí ročne tým, že oddiali schválenie tých užitočných liekov, ktoré sa ešte vyvíjajú a nakoniec budú schválené.

Naozaj teda existuje dôvod byť alrgický na ľudí, ktorí chodia okolo a hovoria: „Ach, ale technológia má aj riziká, nielen výhody.“ Existuje historický záznam ukazujúci, že sme príliš konzervatívni, že mnoho tichých smrtí kvôli regulácii preváži niekoľko viditeľných smrtí kvôli neregulácii. Ak sa *naozaj* chcete hrať na strednú cestu, prečo nepoviete: „Ach, ale technológia má aj výhody, nielen riziká“?

Nuž a toto nie je taký zlý opis pre Zlých Chlapov. (Akurát treba o trochu silnejšie zdôrazňovať, že toto nie sú zlí mutanti ale štandardné ľudské bytosti, konajúce pod vplyvom iného svetonázoru, ktorý im dáva za pravdu; niektorí z nich budú nevyhnutne schopnejší než iní, a táto schopnosť robí veľký rozdiel.) Aj keď sa obzriem naspäť, nemyslím si, že technofília môjho detstva bola príliš mimo ohľadom toho, kto je Zlý Chlap a čo je kľúčová chyba. Ale vždy je *omnoho* ľahšie povedať, čo *nerobiť*, než urobiť to

správne. A jedna z mojich vtedajších základných chýb bolo myslenie, že ak sa snažíte zo všetkých síl vyhnúť všetkému, čo robia Zlí Chlapi, potom ste Dobrý Chlap.

Osobitne škodlivý bol podľa mňa zlý príklad, ktorí dali tí, čo predstierali Hlbokú Múdrosť snahou vytýčiť strednú cestu; usmievať sa rovnako pohŕdajúco na technofilov aj technofóbov, a nazývať oboch nezrelými. Toto naozaj je nesprávna cesta; a v skutočnosti celá predstava o vytýčení strednej cesty vo všeobecnosti je zvyčajne zlá; tá Správna Cesta nie je kompromis medzi niečím, je to čistý prejav svojich vlastných kritérií.

Ale o to ťažšie bolo pre mladého Eliezera zanechať rozhodnutie plnou-parou-vpred, pretože *hocijaký* odklon vyzeral ako pripojenie sa k tým, čo sa hrajú na Hlbokú Múdrosť.

Prvá štrbina v mojej detskej technofílii sa objavila asi v 1997 alebo 1998, keď som si všimol, že moji kolegovia technofili hovoria hlúpe veci o tom, ako by molekulárna nanotechnológia bola ľahko zvládnuteľný problém. (Ako si môžete opäť všimnúť, mladý Eliezer bol ohromne motivovaný svojou schopnosťou nachádzať chyby – dokonca som mal osobnú filozofiu, prečo je niečo takéto dobrý nápad.)

Tá vec s nanotechnológiou by bola na samostatný článok, a možno taký, čo patrí na iný blog. Ale prebiehala debata o molekulárnej nanotechnológii, a o tom, či by útok bol asymetricky jednoduchší než obrana. A našli sa ľudia, ktorí tvrdili, že obrana by bola jednoduchá. V oblasti *nanotechnológie*, doparoma, programovateľnej hmoty, keď to vyzerá, že nedokážeme vyriešiť problém bezpečnosti ani pre počítačové siete, kde môžeme sledovať a kontrolovať každú jednotku a nulu. Ľudia hovorili o nedobytných diamantoidných stenách. Ja som poznamenal, že diamant neodolá jadrovej bombe, že útok porazil obranu už v roku 1945 a nanotechnológia vyzerá, že to pravdepodobne nezmení.

A v čase, keď táto debata skončila, si mladý Eliezer zrejme – zachytený v paľbe argumentov – dokázal uvedomiť, po prvýkrát, že prežitie inteligentného života pochádzajúceho zo Zeme je ohrozené.

Vyzerá to tak zvláštne, pri pohľade naspäť, myslieť si, že existoval čas, keď som si myslel, že sú v budúcnosti ohrozené iba jednotlivé životy. Aký podstatne priateľskejší svet by to bol... hoci vtedy som na to nemyslel. Ani tak som túto možnosť *neodmietal*, ako sa mi skôr *v prvom rade podarilo nikdy ju nevidieť*. Keď táto téma raz naozaj prišla, uvidel som to. Naozaj si nepamätám, ako ten trik fungoval. Existuje dôvod, prečo rozprávam o svojom bývalom ja v tretej osobe.

Môže to znieť, že Eliezer_{1998} bol úplný idiot, ale to by bola príliš pohodlná cesta von, svojím spôsobom; pravda je strašidelnejšia. Eliezer_{1998} bol ostrý Tradičný Racionalista, ako sa požadovalo. Vedel som, že hypotézy majú byť testovateľné, vedel som, že racionalizácia nie je povolená myšlienková operácia, vedel som hrať racionalistické Tabu, bol som posadnutý uvedomovaním si seba samého... celkom som nerozumel pojmu „tajomných odpovedí“... a vôbec som nepoznal Bayesa ani Kahnemana. Aj tak, ostrý Tradičný Racionalista, vysoko nad priemerom... No a čo? Príroda nás neznámkuje podľa rebríčka. Jeden krok mimo Cesty, jedno postrčenie vašich myšlienkových procesov nevhodným vplyvom, môže zrušiť všetky ostatné ochrany.

Jedna z hlavných lekcií, ktoré si odvodzujem z obzerania sa na svoju osobnú históriu je, že sa niet čomu čudovať, že si tam vonku v skutočnom svete veľa ľudí myslí, že „inteligencia nie je všetko“, alebo že racionalistom sa v skutočnom živote nedarí lepšie. Štipka rozumnosti, dokonca ani veľa rozumnosti, neprekročí tú astronomicky vysokú bariéru potrebnú, aby veci naozaj začali *fungovať*.

Nemôžem za svoje nesprávne pochopenie Správnej Cesty obviniť Jerryho Pournellea, svojho otca, ani science fiction vo všeobecnosti. Myslím si, že osobnosť mladého Eliezera vložila veľa selektívnosti do toho, ktoré časti ich učenia prešli. Nie je to tak, že by Pournelle nepovedal: *Keď raz opustíš kolísku, Zem, pravidlá sa zmenia; ak si iba raz lajdácky zapneš svoj skafander, zomrieš.* Povedal to veľakrát. Ale tie slová mi nepripadali naozaj dôležité, pretože to bolo niečo, čo sa stávalo menej dôležitým postavám v románoch – hlavná postava z nejakého dôvodu zvyčajne nezomrela niekde v polovici.

Čo boli sklá, cez ktoré som filtroval tieto učenia? Nádej. Optimizmus. Tešenia sa na svetlejšiu budúcnosť. To bol pre mňa základný význam knihy *O krok ďalej von*, lekcia, ktorú som si vzal v kontraste s osudovosťou a pochmúrnosťou Klubu Sierra. Na jednej strane bola *rozumnosť a nádej*, na druhej *nevedomosť a zúfalstvo*.

Niektorí puberťáci si myslia, že sú nesmrteľní a jazdia na motorkách. Ja som nemal žiadne také ilúzie a dosť som sa vyhýbal šoférovaniu, keď som si uvedomil, ako nebezpečne tie rútiace sa kusy kovu vyzerajú. Ale mal som niečo, čo mi bolo ešte dôležite než môj vlastný život: Budúcnosť. A ja som sa správal, akoby *tá* bola nesmrteľná. Môžeme stratiť životy, ale nie Budúcnosť.

A keď som si uvedomil, že nanotechnológia naozaj *bude* hrozba potenciálne na úrovni vyhynutia?

Mladý Eliezer si pomyslel, vyslovene: „Nebesá, ako je možné, že som si toto nevšimol, keď to malo byť samozrejmé? Musel som byť príliš emocionálne pripútaný k výhodám, ktoré som si od technológie sľuboval; musel som cúvnuť preč od predstavy ľudského vyhynutia."

A potom...

Nevyhlásil som Stop, Rozpusti Sa, Začni Horieť. Nepremyslel som si všetky závery, ktoré som si vyvinul s mojim dovtedajším postojom. Iba sa mi to podarilo *akosi* zapojiť do môjhosveto názoru, s minimom propagovaných zmien. Staré myšlienky a plány boli spochybnené, ale moja myseľ našla dôvody ponechať si ich. Nebolo žiadne systémové zrútenie, nanešťastie.

Najvýznamnejšia zmena bolo moje rozhodnutie, že musíme ísť plnou parou vpred s UI, aby sme ju vyvinuli skôr než nanotechnológiu. Presne tak, ako som to *pôvodne* plánoval urobiť, len teraz z *odlišného dôvodu*.

Takáto je asi väčšina ľudí, nie? Tradičná Rozumnosť nestačila na to, aby to zmenila.

Ale prišiel čas, keď som si naplno uvedomil svoju chybu. Len to chcelo silnejší kopanec do hlavy.

*→

295. Talent na hľadanie chýb

Špirála smrti môjho detstva opisuje základnú zotrvačnosť, ktorá ma odniesla k mojej chybe, afektívnej špirále smrti okolo niečoho, čo Eliezer_{1996} nazýval „inteligencia". Bol som aj technofil, vopred alergický na strach z budúcnosti. A čítal som veľa vedeckej fantastiky, ktorá bola založená na etike osobnosti – kde strach z mimozemšťanov postaví väčšinu ľudstva do role záporných hrdinov, ktorí zle zaobchádzajú s mimozemšťanmi alebo s mysliacimi UI, pretože to „nie sú ľudia".

Toto je časť étosu, ktorý získate z vedeckej fantastiky – definuje vašu skupinu, váš kmeň, primerane široko. Preto som mal mailovú adresu sentience@pobox.com.

Tak sa Eliezer_{1996} rozhodol zostrojiť superinteligenciu, pre dobro ľudstva a všetkého mysliaceho života.

Myslím, že mi spočiatku otázka, či táto superinteligencia bude / môže byť dobrá / zlá vlastne ani nenapadla ako samostatná téma na diskusiu. Iba štandardná intuícia: „Iste by žiadna supermyseľ nebola taká hlúpa, aby premenila celú galaxiu na kancelárske spinky; keby bola taká inteligentná, iste by zároveň vedela, čo je *správne*, omnoho lepšie než by to mohol vedieť človek."

Dokiaľ som sa neprihlásil so svojím poslaním do trashumanistickej mailovej skupiny, a nedostal odpovede v tomto duchu (spamäti):

> Morálka je ľubovoľná – ak hovoríš, že je niečo dobré alebo zlé, nemôžeš mať pravdu alebo sa mýliť. Superinteligencia by si vytvorila svoju vlastnú morálku.
>
> Každý sa v konečnom dôsledku zaujíma o svoj vlastný prospech. Superinteligencia by nebola iná; akurát by zhrabla všetky zdroje.
>
> Ja som osobne človek, preto fandím ľuďom, nie umelým inteligenciám. Nemyslím si, že by sme mali vyvíjať takúto technológiu. Namiesto toho by sme najprv mali vyvinúť technológiu na nahranie ľudí do počítača.
>
> Nikto by nemal vyvíjať UI bez kontrolného systému, ktorý ju sleduje a dáva pozor, aby neurobila nič zlé.

→ http://lesswrong.com/lw/u0/raised_in_technophilia/

Nuž, *toto* je všetko samozrejme zle, pomyslel si Eliezer$_{1996}$ a pokračoval v rozbíjaní argumentov svojich súperov na kúsky. (Väčšinu z toho som už urobil v iných článkoch, a čo zostalo, ponechávam ako cvičenie pre čitateľa.)

Nie je to tak, že by Eliezer$_{1996}$ vedome uvažoval: „Najhlúpejší človek na svete povedal, že slnko svieti, *preto* je tma.“ Ale Eliezer$_{1996}$ bol Tradičný Racionalista; bola mu vštepená metafora vedy ako *férového boja* medzi stranami, ktoré zaujímajú rôzne postoje, akurát bez násilia a iných podobných precvičovaní politických svalov, takže v ideálnom prípade môže vyhrať strana s najlepšími argumentmi.

Je ľahšie povedať, že je argument niekoho iného nesprávny, než zistiť v danej veci fakty správne; a Eliezer$_{1996}$ bol veľmi šikovný v hľadaní chýb. (A to som aj ja. Nemôžete túto nebezpečnú silu zneškodniť tým, že sa odmietnete zaujímať o chyby.) Z pohľadu Eliezera$_{1996}$ to vyzeralo, že jeho zvolená strana *vyhráva boj* – že formuluje lepšie argumenty než jeho súperi – prečo by teda menil stranu?

Preto je písané: „Keďže tento svet obsahuje mnoho tých, ktorých od rozumnosti oddeľuje celá priepasť, začínajúci študenti rozumnosť vyhrávajú argumenty a získavajú prehnaný pohľad na svoje vlastné schopnosti. Byť lepším než druhí je však zbytočné: Život sa neznámkuje podľa rebríčka. Najlepší fyzici v starovekom Grécku nedokázali spočítať dráhu padajúceho jablka. Neexistuje záruka, že ešte aj vaše najťažšie úsilie bude primerané; preto neplytvajte myšlienkami na to, či sa iným darí horšie.“

Nemôžete sa spoliehať na to, že vás z vašich chýb vyargumentuje *niekto* iný; nemôžete sa spoliehať na to, že vás zachráni *niekto* iný; vy a jedine vy máte povinnosť hľadať chyby vo svojich postojoch; ak odložíte toto bremeno, nečakajte, že ho zdvihne niekto iný. A neprekvapilo by ma, keby sa ukázalo, že táto rada väčšine ľudí nepomôže, dokiaľ si osobne neodstrelia vlastnú nohu zatiaľ čo budú sami sebe hovoriť, *pravdivo*: „Je jasné, že túto hádku vyhrávam.“

Dnes sa nepokúšam brať žiadneho človeka ako svojho súpera. To iba vedie k prehnanej sebadôvere. Je to Príroda, komu čelím, ktorá svoje problémy neprispôsobuje vašim schopnostiam, ktorá nemá povinnosť poskytnúť vám férovú šancu na víťazstvo výmenou za usilovnú prácu, ktorú nezaujíma, či ste ten najlepší, aký kedy bol, pokiaľ nie ste *dosť* dobrý.

Ale vráťme sa do roku 1996. Eliezer$_{1996}$ ide podľa základnej intuície: „Superinteligencia bude iste vedieť lepšie než ja, čo je *správne*,“ a spakruky zhadzuje rôzne argumenty vznesené proti jeho postoju. Ako vidíte, bol v tom šikovný. Dokonca mal osobnú filozofiu, prečo je múdre hľadať vo veciach chyby, a tak ďalej.

Nechcem to povedať ako výhovorku, že nikto z tých, čo argumentovali proti Eliezerovi$_{1996}$, mu v skutočnosti neponúkol rozpustenie tohto tajomstva – plnú redukciu morálky, ktorá analyzuje všetky jeho poznávacie procesy debatujúce o „morálke“, sprievodcu krok za krokom po algoritmoch, ktoré spôsobujú, že mu morálka pripadá ako fakt. Považujte to radšej za obvinenie, za mieru úrovne Eliezera$_{1996}$, že by bol potreboval, aby mu dali celé riešenie, aby dostal argument, ktorý *nedokáže* vyvrátiť.

Tých pár prítomných filozofov ho z jeho ťažkostí nevytiahlo. Nie je to tak, že by filozof povedal: „Prepáč, morálka je pochopená, je to vybavená téma v kognitívnej vede a vo filozofii, a tvoj pohľad je jednoducho nesprávny.“ Podstata morálky je vo filozofii stále otvorenou otázkou, debata stále pokračuje. Filozof bude cítiť povinnosť predložiť vám zoznam klasických argumentov za všetky strany; na väčšinu z nich je Eliezer$_{1996}$ dosť inteligentný, aby ich zhodil, a tak dôjde k záveru, že filozofia je pustatina.

Ale počkajte. Ešte to bude horšie.

Nespomínam si presne kedy – mohlo to byť v roku 1997 – ale moje mladšie ja, nazvime ho Eliezer$_{1997}$, sa pustilo do *neúnavného* argumentovania, že vytvorenie superinteligencie je tá správna vec.

*→

→ http://lesswrong.com/lw/u1/a_prodigy_of_refutation/

296. Číre bláznovstvo neoperenej mladosti

> Tu hovorí číre bláznovstvo neoperenej mladosti; nerozvážnosť nevedomosti takej priepastnej, ako je možné iba u jedného z vášho prchavého druhu...
>
> --Gharlane z Eddoru[289]

Kde bolo, tam bolo, pred mnohými rokmi, som navrhol tajomnú odpoveď na tajomnú otázku – ako som už párkrát naznačil. Tá tajomná otázka, na ktorú som navrhol tajomnú odpoveď, však nebola vedomie – alebo skôr, nie len vedomie. Nie, tá zahanbujúcejšia chyba bola, že som prijal tajomný pohľad na morálku.

Vyhýbal som sa diskusii o tomto až doteraz, po sérii o metaetike, pretože som chcel, aby bolo jasné, že to Eliezer$_{1997}$ pochopil *zle*.

Keď sme ho naposledy opustili, Eliezer$_{1997}$ nebol spokojný s argumentovaním v intuitívnom zmysle, že superinteligencia by bola morálna, a pustil sa do *neúnavného* argumentovania, že vytvoriť superinteligenciu je tá správna vec.

Nuž (povedal Eliezer$_{1997}$), začnime otázkou: *Má život v skutočnosti nejaký zmysel?*

„Neviem," odpovedal si Eliezer$_{1997}$ ihneď, s istým tónom blahoželania si k priznaniu svojej nevedomosti na túto tému, kde mnohí iní vyzerali sebaisto.

„Ale," pokračoval...

(Vždy si dávajte pozor, keď za priznaním nevedomosti nasleduje „Ale".)

„Ale, ak predpokladáme, že život nemá zmysel – že užitočnosť všetkých výsledkov sa rovná nule – táto možnosť ruší všetky výpočty *očakávanej* užitočnosti. Môžeme sa teda vždy *správať, ako keby* sme vedeli, že život má zmysel, aj keď nevieme, čo je týmto zmyslom. Ako môžeme tento zmysel nájsť? Vzhľadom na to, že ľudia sa o tomto ešte stále hádajú, je to pravdepodobne problém príliš zložitý na to, aby ho ľudia vyriešili. Potrebujeme teda superinteligenciu, ktorá tento problém vyrieši za nás. A čo sa týka tej možnosti, že *neexistuje* logické zdôvodnenie, prečo niečo uprednostniť pred niečím iným, tak v tom prípade postavenie superinteligencie nie je o nič lepšie ani horšie než hocičo iné. Toto je skutočná možnosť, ale vypadáva z každého pokusu vypočítať očakávaný úžitok – mali by sme ju skrátka ignorovať. Pokiaľ niekto tvrdí, že by superinteligencie vyhubila ľudstvo, buď tým tvrdí, že vyhubenie ľudstva je správna vec (hoci nevidíme žiaden dôvod, prečo by to tak malo byť) alebo tvrdí, že neexistuje nič správne (a v tom prípade jeho argument proti zostrojeniu inteligencie vyvracia sám seba)."

Ech. Tento odsek bolo *naozaj* ťažké napísať. Moje minulé ja je vždy môj najhorší koncentrovaný kryptonit, pretože moje minulé ja, to sú *presne* všetky tie veci, na ktoré si moje moderné ja nainštalovalo alergie, aby ich zablokovalo. Správne sa hovorí, že rodičia robili všetky tie veci, ktoré hovoria svojim deťom, aby ich nerobili, pretože odtiaľ vedia, že je lepšie ich nerobiť; rovnako to platí aj medzi minulými a budúcimi ja.

Nakoľko pomýlený je argument Eliezera$_{1997}$? Nedokážem to ani vymenovať. Viem, že pamäť je nespoľahlivá, rekonštruovaná vždy, keď si spomíname, takže nedôverujem svojej modernej mysli pri skladaní týchto starých kúskov. Nežiadajte ma, aby som si čítal svoje staré zápisky; veľmi to bolí.

Ale zdá sa jasné, že som myslel na úžitok ako na nejaký druh veci, na vnútornú vlastnosť. Takže „život nemá zmysel" zodpovedalo úžitok = 0. Ale samozrejme tento argument funguje rovnako dobre aj s úžitok = 100, takže ak všetko má zmysel, ale všetko má celkom *rovnaký* zmysel, aj to by sa malo vykrátiť... Iste som vtedy nemyslel na funkciu úžitku ako na afinnú štruktúru v preferenciách. Myslel som na „úžitok" ako na absolútnu hladinu vnútornej hodnoty.

Myslel som na *malo by sa* ako na nejaký druh čisto abstraktnej podstaty prinucovateľnosti, to-čo-spôsobuje-že-niečo-robíte; takže samozrejme ľubovoľná myseľ, ktorá odvodí, že by sa niečo *malo*, tým bude viazaná. Preto ten predpoklad, ktorý Eliezerovi$_{1997}$ ani nenapadlo explicitne poznamenať, že logika,

289 Edward Elmer Smith, *Second Stage Lensmen* (Old Earth Books, 1998).

ktorá donúti ľubovoľnú myseľ niečo urobiť, je presne tá istá ako to, na čo myslia ľudia a na čo sa odkazujú, keď vyslovia slovo „správne“...

Ale to už sa snažím vymenovať chyby, a ak ste doteraz sledovali, mali by ste to zvládnuť sami.

Dôležitá stránka celého tohto zlyhania bola, že keďže som dokázal, že *sa neoplatí zaoberať* prípadom „život nemá zmysel“, nemyslel som si, že je potrebné dôkladne definovať „inteligenciu“ alebo „zmysel“. Už som predtým prišiel na chytrý dôvod, prečo sa nepokúšať byť celkom formálny a dôkladný, keď sa snažím definovať „inteligenciu“ (alebo „morálku“) - konkrétne ten chyták, na ktorý sa v minulosti ľudia z UI, filozofi, a moralizátori nachytali, keď si vymysleli vlastné definície, ktorým unikala pointa.

Vyvodzujem nasledujúce ponaučenie: Bez ohľadu na to, aké chytré je zdôvodnenie povoliť vaše štandardy alebo vyhnúť sa nejakej požiadavke dôkladnosti, odstrelíte si tým vlastnú nohu celkom rovnako.

A ďalšie ponaučenie: Bol som šikovný vo vyvracaní. Keby som uplatnil rovnakú úroveň vyvrátenia na základe ľubovoľnej chyby na svoj vlastný postoj, ako som používal na porazenie argumentov vznesených proti mne, potom by som natrafil na logickú medzeru a odmietol tento postoj – keby som *chcel*. Keby som proti nemu mal rovnakú úroveň predsudku ako som mal voči zvyšným postojom v tejto debate.

Lenže toto bolo skôr než som počul o Kahnemanovi, skôr než som počul pojem „motivovaný skepticizmus“, skôr než som si osvojil predstavu presne správneho stavu neistoty, ktorý sumarizuje všetky indície, a skôr než som poznal smrteľnosť otázky: „Môžem tomu veriť?“ pre obľúbené postoje a „Musím tomu veriť?“ pre neobľúbené postoje. Bol som iba Tradičný Racionalista, ktorý myslel na vedecký proces ako na rozhodcu medzi ľuďmi, ktorí zaujali postoje a hádali sa za ne, nech vyhrá najlepšia strana.

Mojou najväčšou chybou nebolo to, že sa mi páčila „inteligencia“, ani žiadne množstvo technofílie a vedeckej fantastiky vyzdvihujúce príbuznosť všetkého vedomia. Iste to nebola moja schopnosť vidieť chyby. Žiadna z týchto vecí by ma *nemohla* odviesť zo správnej cesty, keby som sám od seba žiadal celkovo vyšší štandard dôkladnosti, a odmietol inak zaujať postoj. Alebo keby som aspoň preskúmal môj obľúbený nejasný postoj s rovnakou požiadavkou na dôkladnosť, akú som aplikoval na protiargumenty.

Lenže ja som sa veľmi nezaujímal o pokus vyvrátiť svoj názor, že život má zmysel, keďže moje uvažovanie by bolo vždy dominované prípadmi, kde život má zmysel.

A keď išlo o explóziu inteligencie, myslel som si, že skrátka musím napredovať čo najväčšou rýchlosťou, používajúc tie najlepšie pojmy, aké som mal vtedy k dispozícii, nezastavovať sa a nerušiť všetko, kým budem hľadať dokonalú definíciu, ktorú tak mnoho iných pokazilo...

Nie.

Nie, nepoužívaj tie najlepšie pojmy, aké máš vtedy k dispozícii.

Je to Príroda, kto ťa súdi, a Príroda neprijíma *ani tie najoprávnenejšie výhovorky*. Ak nedosahuješ štandard, zlyháš. Je to také jednoduché. Neexistuje žiaden chytrý argument, prečo si musíš poradiť s tým, čo máš, pretože Príroda nebude tento argument počúvať, neodpustí ti, pretože si mal tak veľa vynikajúcich dôvodov sa ponáhľať.

Všetci vieme, čo sa stalo Donaldovi Rumsfeldovi, keď šiel do vojny s takou armádou, akú mal, namiesto takej armády, akú potreboval.

Možno by Eliezer_{1997} nedokázal vyčarovať správny model len tak zo vzduchu. (Hoci ktovie, čo by sa stalo, keby som sa naozaj pokúsil...) A nebolo by prezieravé, aby celkom prestal myslieť, dokiaľ dôkladnosť zrazu nevyskočí odnikadiaľ.

Ale nebolo ani správne, aby Eliezer_{1997} položil celú svoju váhu na svoj „najlepší odhad“ v neprítomnosti presnosti. Môžete používať nejasné predstavy vo svojich dočasných myšlienkových procesoch, keď hľadáte lepšiu odpoveď, nespokojní so svojimi terajšími nejasnými náznakmi, *a neochotní položiť na ne celú svoju váhu*. Nestaviate však superinteligenciu podľa dočasného chápania. Nie, ani podľa „najlepšieho“ nejasného chápania, aké máte. Toto bola moja chyba – myslieť si, že keď poviem „najlepší odhad“, všetko sa tým ospravedlní. Existoval iba štandard, ktorý som nedosahoval.

Samozrejme Eliezer$_{1997}$ nechcel spomaliť na ceste k explózii inteligencie, keď išlo o tak veľa životov, a o samotné prežitie inteligentného života pochádzajúceho zo Zeme, ak by sme sa dostali do doby nanozbraní pred dobou superinteligencie...

Prírodu nezaujímajú vaše spravodlivé dôvody. Existuje iba astronomicky vysoký štandard potrebný na úspech. Buď ho dosiahnete, alebo ste zlyhali. To je všetko.

> Apokalypsa nemusí byť voči vám férová.
>
> Apokalypsa vám nemusí ponúknuť šancu na úspech.
>
> Výmenou za to, čo ste doteraz urobili.
>
> Zložitosť apokalypsy nie je primeraná vašim schopnostiam.
>
> Cena apokalypsy nie je primeraná vašich prostriedkom.
>
> Ak vás apokalypsa požiada o niečo nerozumné
>
> A vy sa pokúsite trochu zjednávať
>
> (Pretože každý musí tu a tam urobiť kompromis)
>
> Apokalypsa sa nebude snažiť dojednať s vami.

Aha, a ešte to bude horšie.

Ako si Eliezer$_{1997}$ poradil so samozrejmým argumentom, že nemôžete odvodiť „malo by sa“ z čírej logiky, pretože výroky „malo by sa“ možno odvodiť iba z iných výrokov „malo by sa“?

Nuž (všimol si Eliezer$_{1997}$), tento problém má rovnakú štruktúru ako argument, že príčina iba pochádza z inej príčiny, alebo že skutočná vec môže vzniknúť iba z inej skutočnej veci, čím by ste dokázali, že nič neexistuje.

Takže (povedal) existujú tri „ťažké problémy“: Ťažký problém vedomého zážitku, v ktorom vidíme, že kvaliá nemôžu vzniknúť z vypočítateľných procesov; ťažký problém existencie, v ktorom sa pýtame, ako mohla hocijaká existencia vzniknúť z asi ničoho; a ťažký problém morálky, ktorým je ako dostať „malo by sa“.

Tieto problémy asi súvisia. Napríklad kvaliá radosti sú asi jedným z najlepších kandidátov na niečo, čo je žiadúce samo osebe. Možno teda nedokážeme pochopiť ťažký problém morálky bez rozuzlenia ťažkého problému vedomia. Je zrejmé, že tieto problémy sú pre ľudí príliš ťažké – inak by ich niekto bol vyriešil za posledných 2500 rokov od vynálezu filozofie.

Nie je to tak, že by mohli mať zložité riešenia – na to sú príliš jednoduché. Ten problém musí skrátka byť mimo priestoru ľudských pojmov. Keďže vidíme, že vedomie nemôže vzniknúť na žiadnom vypočítateľnom procese, musí zahŕňať novú fyziku – fyziku ktorú používajú naše mozgy, ale nerozumejú jej. Preto potrebujeme superinteligenciu, aby tento problém vyriešila. Dokázateľne to súvisí s kvantovou mechanikou, možno s trochou drobných uzavretých časových slučiek zo všeobecnej relativity; časové paradoxy by mohli mať nejaké rovnaké neredukovateľné vlastnosti, ktoré vedomie asi vyžaduje...

A tak ďalej, až do vyčerpania. Možno začínate vidieť v reťazi mojich článkov na *Overcoming Bias* ten list, ktorý si želám, aby som mohol poslať sám sebe.

Z čoho som si naučil toto: Nemôžete pracovať so zmätkom. Nemôžete dosiahnuť, aby chytré plány fungovali okolo medzier vo vašom chápaní. Nemôžete urobiť ani len „najlepšie odhady“ vo veciach, ktoré vás principiálne mätú, a prirovnať ich k iným mätúcim veciam. Dobre, môžete, ale nebudete to mať správne, dokiaľ sa váš zmätok nerozpustí. Zmätok existuje v mysli, nie v skutočnosti, a skúšať s ním narábať ako s niečím, čo môžete zdvihnúť a posúvať okolo, skončí iba neúmyselnou komédiou.

Podobne, nemôžete prísť s chytrými dôvodmi, prečo na medzerách vo vašom modeli nezáleží. Nemôžete nakresliť hranicu okolo tajomstva, pripevniť naň držadlá, ktoré vám dovolia používať túto Tajomnú Vec aj keď jej v skutočnosti nerozumiete – ako môj pokus vykrátiť možnosť, že život nemá zmysel, zo vzorca očakávaného úžitku. Nemôžete zdvihnúť medzeru a manipulovať ňou.

Ak prázdne miesto na vašej mape ukrýva nášľapnú mínu, potom položiť svoju váhu na toto miesto bude smrteľné, bez ohľadu na to, aké dobré máte výhovorky, prečo to neviete. Ľubovoľná čierna skrinka môže obsahovať pascu, a neexistuje spôsob ako to vedieť, okrem otvorenia tejto čiernej skrinky a pohľadu dovnútra. Ak prídete s nejakým spravodlivým zdôvodnením, prečo sa musíte ponáhľať vpred s najlepším porozumením, aké máte – pasca sa spustí.

> Až keď poznáš pravidlá,
>
> Potom si uvedomíš, *prečo* si sa potreboval učiť;
>
> Čo by sa bolo stalo inak,
>
> Ako *veľa* si potreboval vedieť.

Iba poznanie dokáže predpovedať cenu nevedomosti. Starovekí alchymisti nemali žiadnu logickú možnosť poznať presné dôvody, prečo je také ťažké premeniť olovo na zlato. Preto sa otrávili a zomreli. Prírode je to jedno.

Ale prišiel čas, keď mi začalo svitať.

*→

297. Ten tenký tón nesúladu

Keď sme naposledy opustili Eliezera_{1997}, bol presvedčený, že ľubovoľná superinteligencia by automaticky robila to, čo je „správne“, a vlastne by tomu rozumela lepšie než by sme mohli my; hoci, ako skromne priznával, nerozumel konečnej podstate morálky. Alebo skôr, po uplynutí nejakých debát si Eliezer_{1997} vyvinul prepracovaný argument, o ktorom s nehou tvrdil, že je „formálny“, že by sme mohli vždy vychádzať z predpokladu, že život má zmysel; takže by prípady, keď by superinteligencie necítili nutkanie robiť nič konkrétne, vypadli z našich úvah. (Chybou bolo nedomyslené a nezdôvodnené stotožnenie „všeobecne presvedčivého argumentu“ so „správnym“.)

Zatiaľ je teda mladý Eliezer na dobrej ceste pridať sa k „bystrým ľuďom, ktorí sú hlúpi, pretože sú šikovní v obhajovaní názorov, ktoré získali nešikovným spôsobom“. Všetka jeho oddanosť „rozumnosti“ ho nezachránila pred touto chybou, a mohli by ste byť v pokušení dôjsť k záveru, že usilovať sa o rozumnosť je zbytočné.

Lenže hoci si mnoho ľudí kope jamu pod sebou, nie každému sa podarí vyškriabať sa z nej opäť von.

A z tohto som si odvodil ponaučenie: Že to celé začalo...

...malou, maličkou otázkou; jedným neladiacim tónom; jednou drobnou osamelou myšlienkou...

Náš príbeh začína o tri roky neskôr u Eliezera_{2000}, ktorý vo väčšine vlastností pripomína svoje ja z roku 1997. Práve si myslí, že dokázal, že postaviť superinteligenciu je tá správna vec, ak vôbec nejaká správna vec existuje. Z čoho vyplýva, že neexistuje *zdôvodniteľný* konflikt záujmu ohľadom explózie inteligencie medzi národmi a ľuďmi Zeme.

Toto je pre Eliezera_{2000} dôležitý záver, pretože považuje predstavu bojov o explóziu inteligencie za *neznesiteľnej* hlúpu. (Niečo ako predstava, že Boh zasahuje do bojov medzi kmeňmi hašterivých barbarov, akurát naopak.) Predstava Eliezera_{2000} o sebe samom mu nedovoľuje – ani by to *nechcel* – pokrčiť plecami a povedať: „Nuž, naši sa sem dostali ako prví, takže si vezmeme všetky banány skôr než sa k nim dostane niekto iný.“ Je to príliš bolestivá myšlienka.

A predsa mu potom napadne toto:

> Možno by niektorí ľudia dali prednosť tomu, aby UI robila nejaké konkrétne veci, povedzme aby ich nezabila, aj v prípade, že život nemá zmysel?

Jeho okamžite nasledujúca myšlienka je samozrejmá, vzhľadom na jeho predpoklady:

→ http://lesswrong.com/lw/u2/the_sheer_folly_of_callow_youth/

> V tom prípade, že život nemá zmysel, nič nie je „správne“; preto by nebolo zvlášť správne rešpektovať preferencie ľudí v tejto veci.

Toto je zrejmé vyhnutie sa. Eliezer$_{2000}$ však o sebe nerozmýšľa ako o darebákovi. Nechodí okolo so slovami: „Ktorým guľkám sa vyhnem dnes?“ Rozmýšľa o sebe ako o dôkladom racionalistovi, ktorý húževnato nasleduje, kam ho skúmanie zavedie. Neskôr sa obzrie a uvidí mnoho skúmania, ktorému sa jeho myseľ akosi vyhla – ale to nie je jeho *terajšie zmýšľanie o sebe*.

Takže Eliezer$_{2000}$ sa len tak *nechytí* samozrejmej výhovorky. Rozmýšľa ďalej.

> Lenže ak si ľudia myslia, že majú preferencie aj v tom prípade, že život nemá zmysel, potom majú motív nesúhlasiť s mojím projektom explózie inteligencie a dať prednosť projektu, ktorý rešpektuje ich želanie aj v prípade, že život nemá zmysel. Toto vytvára aktuálny konflikt záujmov ohľadom explózie inteligencie a bráni tomu, aby sa stali správne veci v tom hlavnom prípade, ak život má zmysel.

Nuž, existuje *veľa* výhovoriek, ktoré Eliezer$_{2000}$ potenciálne mohol použiť, aby vyhodil tento problém z okna. Viem, pretože som *počul* hromadu výhovoriek zavrhujúcich Priateľskú UI. „Toto je príliš ťažký problém na vyriešenie“ je jedna, ktorú dostávam od rádoby VUI expertov, ktorí si predstavujú, že sú dosť bystrí na to, aby vytvorili skutočnú umelú inteligenciu, ale nie dosť bystrí na to, aby vyriešili naozaj zložitý problém ako je Priateľská UI. Alebo „znepokojovať sa nad touto možnosťou by bolo zlým vynaložením prostriedkov, zabúdaš na to, aké neuveriteľne súrne je vytvoriť UI skôr než sa ľudstvo zlikviduje samo – musíš pracovať s tým, čo máš“, čo hovoria ľudia, ktorých tento problém jednoducho nezaujíma.

Lenže Eliezer$_{2000}$ je *perfekcionista*. Nie je dokonalý, samozrejme, a neprikladá cnosti *presnosti* takú dôležitosť ako ja, ale je to celkom isto *perfekcionista*. Predstava metaetiky, ktorú Eliezer$_{2000}$ zastáva, podľa ktorej superinteligencie vedia, čo je správne, lepšie než my, doteraz napohľad zabalila *všetky* problémy spravodlivosti a morálny do vzduchotesného obalu.

Táto nová výhrada akoby prepichla malú dierku do tohto vzduchotesného obalu. Treba to zaplátať. Keby ste mali niečo, čo je dokonalé, boli by ste naozaj ochotní nechať jednu malú možnosť, aby to skompromitovala?

Takže Eliezer$_{2000}$ sa tejto téme ani *nechce* vyhnúť; chce zaplátať tento problém a obnoviť dokonalosť. Ako môže zdôvodniť strávenie času? Myslením si vecí ako:

> A čo Brian Atkins? [Brian Atkins je zakladajúci sponzor Výskumného ústavu strojovej inteligencie.] On by pravdepodobne radšej nezomrel, aj keby život nemal zmysel. On teraz financuje MIRI; ja nechcem pošpiniť etiku našej spolupráce.

Zmýšľanie Eliezera$_{2000}$ sa nedá preložiť veľmi dobre – angličtina na to nemá jednoduchý popis, ani žiada iná kultúra, ktorú poznám. Možno pasáž v Starom Zákone: „Nebudeš variť kozľa v mlieku jeho matky.“ Niekto, kto vám pomôže z altruizmu, by nemal ľutovať, že vám pomohol; dlžíte mu, ani nie vernosť, ale skôr to, že naozaj robí to, čo si myslí, že robí tým, že vám pomáha.

No ale ako by na to Brian Atkins došiel, keby som mu to nepovedal? Eliezer$_{2000}$ si takéto veci *nemyslí*, jedine ak v úvodzovkách ako niečo, čo by si v takejto situácii pomyslel darebák. Eliezer$_{2000}$ má aj pripravenú štandardnú proti-myšlienku, stráž proti pokušeniam nečestnosti – argument, ktorý zdôvodňuje poctivosť v jazyku očakávaného úžitku, nie iba osobnej lásky k osobnej cnosti:

> Ľudia nie sú dokonalí podvodníci; je pravdepodobné, že by ma niekto odhalil. Alebo čo ak niekto vynájde skutočné detektory lži pred Singularitou, niekedy počas nasledujúcich tridsiatich rokov? Nedokázal by som potom prejsť testom detektorom lži.

Eliezer$_{2000}$ žije podľa pravidla, že by ste mali byť vždy pripravení odvysielať svoje myšlienky hocikedy celému svetu, bez zahanbenia. Inak je jasné, že ste padli z milosti: buď si myslíte niečo, čo by ste si nemali myslieť, alebo sa hanbíte za niečo, za čo by ste sa nemali hanbiť.

(V dnešnej dobe už nepresadzujem takýto extrémny pohľad, hlavne z dôvodov teórie zábavy. Vidím priestor pre pokračujúcu spoločenskú súťaž medzi inteligentnými formami života, prinajmenšom pokiaľ siaha moja krátkodobá vízia. Pripúšťam, že v dnešnej dobe môže byť celkom správne, aby ľudia mali svoje ja; ako to vyjadril John McCarthy: „Keby mal každý žiť pre druhých po celý čas, život by bol ako zástup mravcov, ktorí jeden druhého nasledujú v kruhu.“ Ak máte mať svoje ja, rovnako dobre môžete mať aj tajomstvá, a možno aj konšpirácie. Ale stále sa snažím riadiť princípom byť schopný prejsť testom budúceho detektora lži, pred niekým, kto je tiež ochotný ísť pod detektor lži, ak je to profesionálna záležitosť. Teória zábavy potrebuje výnimku zdravého rozumu pre menežment rizika globálnej katastrofy.)

Ešte aj keď berieme poctivosť za danú, stále sú iné výhovorky, ktoré Eliezer$_{2000}$ mohol použiť na spláchnutie tejto otázky do záchoda. „Svet na to nemá čas“ alebo „To sa nedá vyriešiť“ by stále fungovalo. Ale Eliezer$_{2000}$ *nevie*, že teto problém, problém „záložnej“ morálky, bude zvlášť ťažký alebo časovo náročný. Iba teraz o celej tej veci začal rozmýšľať.

A tak Eliezer$_{2000}$ *začne* skutočne zvažovať otázku: Predpokladajme, že „život nemá zmysel“ (že superinteligencie *nedokážu* vytvoriť svoju motiváciu z čistej logiky), ako by ste potom špecifikovali *záložnú* morálku? Zosyntetizovali by ste ju, vpísali by ste ju do UI?

V tomto bode je toho veľa, čo Eliezer$_{2000}$ nevie. Ale *rozmýšľal* o sebazdokonaľujúcej UI už tri roky, a Tradičným Racionalistom bol ešte dlhšie. Existujú techniky rozumnosti, ktoré *praktizoval*, metodologické poistky, ktoré si už vytvoril. Už vie veľa na to, aby si myslel, že jediné, čo UI potrebuje, je Jeden Veľký Morálny Princíp. Eliezer$_{2000}$ už vie, že je múdrejšie myslieť technologicky než politicky. Už pozná porekadlo, že programátori UI by mali myslieť v kóde, používať pojmy, ktoré možno vpísať do počítača. Eliezer$_{2000}$ už má predstavu, že existuje niečo, čo sa volá „technické myslenie“ a je to dobré, hoci ešte nesformuloval bayesovský pohľad na to. A už si dávno všimol, že sugestívne nazvané symboly LISP-u naozaj nič neznamenajú, a tak ďalej. Tieto zákazy mu zabránia padnúť do niektorej z úvodných pascí, do tých, ktoré som videl, ako pohltili iných nováčikov pri ich prvých krokoch k problému Priateľskej UI... hoci technicky bol toto môj *druhý* krok; pri tom prvom som poriadne a dôkladne zlyhal.

Ale nakoniec to došlo k tomuto: Po prvýkrát sa Eliezer$_{2000}$ pokúša rozmýšľať technicky nad vpísaním morálky do UI, bez únikových dvierok tajomnej esencie správnosti.

To je jediné, na čom nakoniec záleží. Jeho predchádzajúce filozofovanie nestačilo na to, aby donútilo jeho mozog čeliť podrobnostiam. Tento nový štandard je dosť prísny na to, aby si vyžadoval skutočnú prácu. Morálka postupne začne vyzerať menej tajomne – Eliezer$_{2000}$ začína rozmýšľať o *vnútre* čiernej skrinky.

Jeho *dôvody*, prečo sa do tohto konania pustil – na tých vôbec nezáleží.

Aha, je to ponaučenie z jeho perfekcionizmu. Je to ponaučenie z tej časti, ako si Eliezer$_{2000}$ pôvodne myslel, že toto je drobná chyba, a mohol ju pustiť z hlavy, keby mal taký impulz.

Ale nakoniec ide reťaz príčiny a následkov takto: Eliezer$_{2000}$ riešil veci podrobnejšie, preto sa praxou zlepšil. Činy odtieňujú zdôvodnenia. Ak váš argument zhodou okolností zdôvodňuje nepracovanie na podrobnostiach, ako Eliezer$_{1996}$, potom sa nezlepšíte v rozmýšľaní o probléme. Ak váš argument žiada, aby ste pracovali na podrobnostiach, potom máte *príležitosť* začať zhromažďovať odbornosť.

To bola jediná voľba, na ktorej na konci záležalo – nie *dôvody*, prečo niečo robiť.

Toto všetko hovorím, ako ste mohli uhádnuť, kvôli rádoby UI expertom, na ktorých občas narazím, ktorí majú svoje vlastné chytré dôvody prečo nerozmýšľať o probléme Priateľskej UI. Naše chytré dôvody, prečo robiť to, čo robíme, znamenajú pre Prírodu omnoho menej než znamenajú pre nás a pre

našich priateľov. Ak vaše činy nevyzerajú dobre, keď sa zbavia všetkých svojich zdôvodnení a predložia sa iba ako holé fakty... tak by ste ich možno mali znovu preskúmať.

Usilovná práca človeka vždy nezachráni. Existuje aj také niečo ako nedostatok schopnosti. Ale aj tak, ak to neskúsite, ak to neskúsite dosť silno, nedostanete šancu sadnúť si k stolu s vysokými stávkami – a už vôbec nie tú schopnosť. Toto je pre vás príčina a následok.

Navyše, na perfekcionizme naozaj záleží. Koniec sveta nemusí vždy prísť s trúbkami a hromom a najvyššou prioritou vo vašej mailovej schránke. Niekedy sa vám zdrvujúca pravda po prvýkrát ohlási ako malá, maličká otázka; jeden neladiaci tón; jedna drobná osamelá myšlienka, ktorú by ste mohli zamietnuť jediným ľahkým pohybom...

...a tak počas nasledujúcich rokov, začalo tomuto minulému Eliezerovi postupne svitať. Toto slnko vychádzalo pomalšie než by mohlo.

*→

298. Bojovať zákopovú vojnu proti pravde

Keď sme naposledy opustili Eliezera_{2000}, práve začínal skúmať otázku, ako vpísať morálku do UI. Jeho dôvody, prečo to robil, sú celkom nepodstatné, nanajvýš ako zhodou okolností historická ukážka dôležitosti perfekcionizmu. Ak niečo precvičujete, môžete sa v tom zdokonaliť; ak niečo skúmate, môžete to zistiť; jediná vec, na ktorej záleží, je že Eliezer_{2000} v skutočnosti naplno sústredil svoju energiu na technické rozmýšľanie o morálke UI; namiesto ako predtým, na hľadanie zdôvodnenia prečo svoj čas takto netráviť. Na konci, toto bolo jediné, na čom záležalo.

Ale ako začína náš príbeh – ako sa obloha rozjasňuje do sivej a za obzorom vykukne vrchol slnka – Eliezer_{2001} si ešte nepripustil, že Eliezer_{1997} bol *pomýlený* v dôležitom zmysle. On iba *vylepšuje* stratégiu Eliezera_{1997} tým, že pridáva *záložný* plán pre „tú nepravdepodobnú možnosť, že by sa ukázalo, že život nemá zmysel“...

...čo znamená, že Eliezer_{2001} má teraz ústupovú líniu od svojej chyby.

Nemyslím tým len, že Eliezer_{2001} môže povedať: „Priateľská UI je záložný plán“ namiesto výkriku: „JOJ!“

Myslím tým, že Eliezer_{2001} naozaj *má* záložný plán. Ak Eliezer_{2001} začne pochybovať o svojej metaetike z roku 1997, Singularita má záložnú stratégiu, menovite Priateľskú UI. Eliezer_{2001} môže spochybňovať svoju metaetiku a nebude to signalizovať koniec sveta.

A tento prechod je hladký; môže pripustiť šancu 10 %, že sa predtým mýlil, potom šancu 20 %. Nemusí vykašľať celú svoju chybu v jednom veľkom kuse.

Ak si myslíte, že to znie, že je Eliezer_{2001} príliš pomalý, ja celkom súhlasím.

Stratégie $\text{Eliezera}_{1996\text{-}2000}$ boli vytvorené v úplnej neprítomnosti „Priateľskej UI“ ako námetu na úvahu. Celá myšlienka bola získať superinteligenciu, *hocijakú* superinteligenciu, tak rýchlo, ako sa len dá – polievka z kúskov kódu, ad-hoc heuristiky, evolučné programovanie, open source, hocičo čo vyzerá, že by mohlo fungovať – najlepšie všetky prístupy naraz v Projekte Manhattan. („Všetci rodičia robili veci, o ktorých hovoria svojim deťom, aby ich nerobili. Odtiaľ vedia, že sa to robiť nemá.“[290]) Pridanie jedného prístupu navyše nemôže *uškodiť*.

Jeho postoje k technologickému pokroku už boli sformované – alebo presnejšie, zachované z technofílie absorbovanej v detstve – okolo predpokladu, že hocijaký/každý pohyb smerom k superinteligencii je čisté dobro bez náznaku nebezpečenstva.

→ http://lesswrong.com/lw/u7/that_tiny_note_of_discord/

290 John Moore, *Slay and Rescue* (Xlibris Corp, 2000).

Keď sa obzriem, čo *mal* Eliezer_{2001} v tomto bode urobiť, bolo vyhlásiť udalosť HMC (Halt, Melt, and Catch Fire) – Stoj, Roztop Sa, Začni Horieť. Jeden zo základných predpokladov, na ktorom bolo všetko ostatné postavené, sa ukázal ako nesprávny. Toto si vyžaduje zabrzdiť myšlienky až do úplného zastavenia: zložte svoju váhu zo všetkých názorov postavených na nesprávnom predpoklade, a snažte sa čo najviac premyslieť všetko od základu. Toto je umenie, o ktorom potrebujem napísať viac – podobá sa to na kŕčovité úsilie potrebné na vážne domáce upratovanie po tom, čo si dospelý veriaci po prvýkrát všimne, že Boh neexistuje.

Ale čo Eliezer_{2001} robil naozaj, bolo opakovanie si svojich predchádzajúcich technofilných argumentov, prečo je ťažké zakázať alebo vládou kontrolovať nové technológie – štandardné argumenty proti „zrieknutiu sa“.

Ešte aj môjmu modernému ja sa zdá, že všetky tieto hrozné dôsledky, o ktorých technofili hovoria, že nasledujú po rôznych druhoch vládnej regulácie, sú viacmenej správne – je omnoho ľahšie povedať, že niekto to robí zle, než povedať, ako je to správne. Môj moderný pohľad sa neposunul k myšlienke, že technofili sa mýlia ohľadom nevýhod technofóbie; ale mám sklon byť omnoho ústretovejší voči tomu, čo technofóbovia hovoria o nevýhodách technofílie. Čo predchádzajúci Eliezeri povedali o ťažkostiach napríklad, aby vláda urobila niečo zmysluplné ohľadom Priateľskej UI, to stále vyzerá celkom pravdivo. Akurát že mi mnohé jeho nádeje vo vedu, alebo súkromný priemysel, atď., teraz pripadajú rovnako pomýlené.

Nezachádzajme teraz do podrobností technovratkého hľadiska. Eliezer_{2001} práve vyhodil veľký základný predpoklad – že UI, na rozdiel od iných technológií, nemôže byť nebezpečná – von oknom. Intuitívne by ste očakávali, že by to malo mať nejaký veľký účinok na jeho stratégiu.

Nuž, Eliezer_{2001} sa aspoň vzdal svojej predstavy z roku 1999 o open-source Projekte Manhattan UI používajúcej sebamodifikujúcu zmes heuristík, ale celkovo...

Celkovo, predtým chcel vtrhnúť dnu s vytaseným koltom, okamžite použiť svoj najlepší vtedajší nápad; a potom stále chcel vtrhnúť dnu s vytaseným koltom. Nepovedal: „Neviem, ako toto urobiť.“ Nepovedal: „Potrebujem viac poznania.“ Nepovedal: „Tento projekt nie je pripravený na to, aby sme začali programovať.“ Stále to bolo: „Hodinky tikajú, musíme sa hýbať! MIRI začne programovať akonáhle dostaneme dosť peňazí!“

Predtým sa chcel sústrediť na čo najviac vedeckého úsilia s plným zdieľaním informácií, a potom stále myslel v týmto pojmoch. Vedecké tajomstvo = záporný hrdina, otvorenosť = kladný hrdina. (Eliezer_{2001} nečítal o Projekte Manhattan a nevedel o podobnej hádke, ktorú mali Leo Szilard a Enrico Fermi.)

Toto je ten problém s premenou jedného veľkého „Joj!“ na plynulý prechod meniacej sa pravdepodobnosti. Znamená to, že nie je jeden moment prelievania sĺz – viditeľný obrovský dopad – ktorý by naznačil, že možno treba rovnako obrovské zmeny.

Namiesto toho sú tu všetky tieto malé zmeny v názoroch... ktoré vám dávajú šancu opraviť *argumenty* pre svoje stratégie; o kúsok posunúť zdôvodnenie, ale ponechať „základnú myšlienku“ na svojom mieste. Malé šoky, ktoré systém dokáž absorbovať bez prasknutia, pretože zakaždým dostane šancu vrátiť sa a opraviť sa. Akurát že v oblasti rozumnosti, trhliny = dobré, opravy = zlé. V umení rozumnosti je omnoho efektívnejšie pripustiť jednu obrovskú chybu, než pripustiť omnoho malých chybičiek.

Myslím si, že ľudia majú nejaký inštinkt chrániť svoje pôvodné stratégie a plány, aby ustavične neskákali dokola a neplytvali zdrojmi; a samozrejme inštinkt chrániť každý postoj, za ktorý sme verejne argumentovali, aby sme neutrpeli poníženie, že sme sa mýlili. A hoci sa mladý Eliezer usiloval toľké roky o rozumnosť, nie je voči týmto impulzom imúnny; jemným vánkom ovplyvňujú jeho myšlienky, a toto že žiaľ viac než dostatočná škoda.

Ešte v roku 2002 si dávnejší Eliezer nie je celkom *istý*, či by plán Eliezera_{1997} *predsa len nemohol* fungovať. *Mohlo* to dopadnúť dobre. Človek nikdy nevie, však?

Ale prišiel čas, keď to celé s rachotom spadlo.

*→

299. Moje naturalistické prebudenie

Vo včerajšej časti Eliezer_{2001} bojoval zákopovú vojnu proti pravde. Iba postupne posúval svoje názory, pripúšťal rastúcu pravdepodobnosť iného scenára, ale nikdy na rovinu nepovedal: „Mýlil som sa.“ Opravoval svoje stratégie, keď boli spochybnené, hľadal nové zdôvodnenia, pre ten istý plán, aký si vytýčil predtým.

(O čom sa preto hovorí: „Vyvarujte sa bojovania zákopovej vojny proti indíciám, neochotne ustupujúc o každý krok územia iba keď ste donútení, cítiac sa oklamaní. Vzdajte sa pravde tak rýchlo, ako len dokážete. Urobte to v okamihu, keď si uvedomíte, že kladiete odpor; v okamihu, ked vidíte, z ktorých končín veje vietor indícií proti vám.“)

Pamäť slabne a ja ťažko zvládam obzerať sa späť na tieto časy – nie, vážne, *neznášam* čítať svoje staré písanie. Už ma raz opravili v mojich spomienkach tí, ktorí tam vtedy boli. A tak, hoci si pamätám tie dôležité udalosti, nie som si celkom istý, v akom *poradí* sa stali, a tobôž v ktorý rok.

Ale keby som mal vybrať okamih, keď sa moje bláznovstvo zlomilo, vybral by som okamih, keď som po prvýkrát pochopil, celkom vo všeobecnosti, pojem optimalizačného procesu. To bol bod, keď som sa prvýkrát obzrel späť a povedal: „Bol som blázon.“

Predtým, v roku 2002 som trochu písal o evolučnej psychológii ľudskej všeobecnej inteligencie – hoci som si v tom čase myslel, že píšem o UI; v tom bode som si myslel, že som proti antropomorfnej inteligencii, ale stále som pozeral na ľudský mozog ako na inšpiráciu. (Spomínaný článok je „Úrovne organizácie vo všeobecnej inteligencii“, vyžiadaná kapitola pre zborník *Všeobecná umelá inteligencia*,[291] ktorý nakoniec vyšiel v tlači v roku 2007.)

Takže som rozmýšľal (a písal) o tom, ako sa prirodzenému výberu podarilo vypľuť ľudskú inteligenciu; videl som medzi nimi *rozdiel*, sleposť prirodzeného výberu a prezieravosť inteligentného predvídania, uvažovanie pomocou simulácie verzus skúšanie všetkého naživo, abstraktné verzus konkrétne myslenie. A predsa to bol prirodzený výber, čo vytvoril ľudskú inteligenciu, takže naše mozgy, hoci nie naše myšlienky, sú vytvorené celkom podľa rukopisu prirodzeného výberu.

Do dnešného dňa mi toto pripadá ako pomerne zdrvujúci vhľad, a tak ma vytáča, keď ľudia hádžu prirodzený výber a inteligenciou poháňané procesy do jedného vreca ako „evolučné“. Naozaj sú v mnohých veciach takmer absolútne odlišné – aj keď existujú spoločné veci, pomocou ktorých by sa dali opísať, ako konsekvencializmus alebo všeobecnosť vo viacerých oblastiach.

Ale to, že Eliezer_{2002} rozmýšľa v pojmoch *protikladu* medzi evolúciou a inteligenciou, vám niečo povie o hraniciach jeho vízie – ako keď niekto rozmýšľa o politike ako o protiklade medzi konzervatívnym a liberálnym postojom, alebo niekto, kto rozmýšľa o ovocí ako o protiklade medzi jablkami a jahodami.

Potom, čo bol náčrt „Úrovní organizácie“ uverejnený na internete, Emil Gilliam poukázal na to, že môj pohľad na UI sa pomerne podobá na môj pohľad na inteligenciu. Eliezer_{2002} samozrejme neschvaľuje zostrojenie UI na obraz ľudskej mysle; Eliezer_{2002} dobre vie, že ľudská myseľ je iba zlátanina vypľutá prirodzeným výberom. Ale Eliezer_{2002} opísal tieto úrovne organizácie v ľudskom myslení, a nenavrhol používanie iných úrovní organizácie v UI. Emil Gilliam sa opýtal, či si nemyslím, že lovím príliš blízko ľudských hraníc. Nazval som alternatívu „Celkom Mimozemský Dizajn Mysle“ a odpovedal som, že CMDM je asi pre ľudských inžinierov príliš ťažké vytvoriť, aj keď je to teoreticky možné, pretože by sme niečomu takému mimozemskému nedokázali porozumieť, kým by sme to skladali.

→ http://lesswrong.com/lw/u8/fighting_a_rearguard_action_against_the_truth/

291 Ben Goertzel and Cassio Pennachin, eds., *Artificial General Intelligence*, Cognitive Technologies (Berlin: Springer, 2007), doi:10.1007/978-3-540-68677-4.

Neviem, či Eliezer$_{2002}$ vymyslel túto odpoveď sám, alebo či ju prečítal niekde inde. Treba však povedať, že som odvtedy počul túto výhovorku mnohokrát. V skutočnosti, ak niečomu naozaj rozumiete, zvyčajne to viete prestaviť do takmer ľubovoľného tvaru, ponechajúc v tom nejaký štrukturálny základ; ale ak nerozumiete lietaniu, predpokladáte, že lietajúci stroj potrebuje perie, pretože sa nedokážete v predstavách odtrhnúť od analógie s vtákom.

Takže Eliezer$_{2002}$ je stále v istom zmysle pripútaný ku kvázi-ľudským dizajnom mysle – predstavuje si, že ich vylepší, ale ľudská *architektúra* je stále v istom zmysle jeho východiskom.

Čo konečne pretrhne toto puto?

Je to zahanbujúce priznanie: Prišlo to zo sci-fi príbehu, ktorý som sa pokúšal napísať. (Nie, nemôžete ho vidieť; nie je hotový.) Ten príbeh obsahoval nie-kognitívny a nie-evolučný optimalizačný proces; niečo ako Pumpa na výsledky. Nie inteligencia, ale cez-časový fyzikálny jav – čiže, predstavoval som si to ako fyzikálny jav – ktorý úzko ohraničoval priestor možných výsledkov. (Viac vám nemôžem povedať; to by som prezradil pointu, keby som tento príbeh náhodou raz dokončil. Pozrite si skrátka článok o Pumpách na výsledky.) Bol to „iba príbeh“, preto som sa voľne pohrával s touto myšlienkou a logicky som ju rozoberal: bolo dané, že sa stane C, preto bolo dané, že sa (v minulosti) stane B, a preto bolo dané, že sa stane A (ktoré viedlo k B).

Kresliť čiaru cez jeden bod sa vo všeobecnosti považuje za nebezpečné. Dva body tvoria protiklad; predstavujete si jeden ako opak druhého. Ale keď máte tri rôzne body – vtedy ste nútení zobudiť sa a zovšeobecniť.

Teraz som mal tri body: Ľudská inteligencia, prirodzený výber, a moje fiktívne zariadenie.

A tak to bol bod, v ktorom som zovšeobecnil pojem optimalizačného procesu, procesu, ktorý stláča budúcnosť do úzkej oblasti možného.

Toto môže vyzerať ako zrejmá vec, ak ste celý čas sledovali *Overcoming Bias*; ale ak sa pozriete na zbierku 71 definícií inteligencie Shanea Legga, uvidíte, že „stláčanie budúcnosti do ohraničenej oblasti“ je menej samozrejmá odpoveď než sa zdá.

Mnohé z definícií „inteligencie“ od výskumníkov UI hovoria o „riešení problémov“ alebo „dosahovaní cieľov“. Ale prinajmenšom z pohľadu minulých Eliezerov je to až spätný pohľad, ktorý z toho robí to isté ako „stláčanie budúcnosti“.

Cieľ je myšlienkový objekt; elektróny nemajú ciele, ani neriešia problémy. Keď si človek predstavuje cieľ, predstavuje si činiteľa obdareného chcením – je to stále jazyk empatie.

Môžete vysloviť názor, že inteligencia je o „dosahovaní cieľov“ - a potom sa otočiť naopak a začať sa hádať, či sú niektoré „ciele“ lepšie než iné – alebo začať hovoriť o múdrosti potrebnej na to, aby ste posudzovali samotné ciele – alebo hovoriť o systéme, ktorý vedome upravuje svoje ciele – alebo hovoriť o tom, že potrebujeme slobodnú vôľu, aby *vyberala* plány, ktoré dosahujú ciele – alebo hovoriť o tom, ako si UI uvedomí, že jej ciele nie sú to, o čo ju naozaj chceli požiadať jej programátori. Ak si predstavíte niečo, čo stláča budúcnosť do úzkej oblasti možného, ako je Pumpa na výsledky, tieto napohľad zmysluplné výroky sa akosi nepreložia.

Takže prinajmenšom pre mňa, prekuknúť slovo „myseľ“ ako fyzikálny proces, ktorý by svojím prirodzeným fungovaním, riadením sa fyzikálnymi zákonmi, nakoniec natlačil svoju budúcnosť do úzkej oblasti, bolo naturalistickým osvietením oproti predstave činiteľa, ktorý sa snaží dosiahnuť svoje ciele.

Bolo to ako vypadnúť z hlbokej jamy, dopadnúť do bežného sveta, kognitívne napäté tlaky sa uvoľnili do nenútenej jednoduchosti, zmätok sa rozplynul na dym a odplával preč. Videl som *prácu*, ktorú *robí* inteligencia; *bystrosť* už nebola vlastnosť, ale stroj. Ako uzol v čase, ktorý odráža vonkajšiu časť vesmíru vo vnútornej časti, a tým ju naviguje. Dokonca som uvidel, vo svetle tohto osvietenia, že myseľ musí produkovať odpadové teplo, aby dodržiavala zákony termodynamiky.

Predtým Eliezer$_{2001}$ hovoril o Priateľskej UI ako o niečom, čo by ste mali urobiť len pre istotu – keby ste nevedeli, či dizajn UI X bude Priateľský, potom by ste mali uprednostniť dizajn UI Y, o ktorom

viete, že bude Priateľský. Ale Eliezer_{2001} si nemyslel, že *vie*, či je *naozaj* možná superinteligencia, ktorá premieňa svoj svetelný kužeľ na kancelárske spinky.

Teraz to však *vidím* – rytmus optimalizačného procesu, vlievajúce sa zmyslové informácie, vylievajúce sa pohybové pokyny, navigujúce budúcnosť. Uprostred je model, ktorý spája možné akcie s možnými výsledkami, a funkcia úžitku pre tieto výsledky. Vložte príslušnú funkciu úžitku, a výsledkom bude optimalizátor, ktorý bude navigovať budúcnosť kamkoľvek.

Až do tohto bodu som si nikdy celkom nepripustil, že systém dizajnu cieľov UI Eliezera_{1997} by definitívne, jednoznačne, zbytočne vyhubil celé ľudstvo. Teraz som sa však obzrel naspäť a konečne uvidel, *čo môj starý dizajn robil naozaj*, do tej miery, do akej bol vôbec koherentný. Približne povedané, premenil by celý svoj budúci svetelný kužeľ na všeobecné nástroje – počítače, na ktorých by nebežali žiadne programy, uloženú energiu, ktorá by nemala použitie...

...ako je možné, že ja, šikovný a skúsený racionalista, ako je možné, že som si dokázal nevšimnúť niečo také zrejmé, po celých šesť prekliatych rokov?

Toto bol bod, keď som sa zobudil s jasnou hlavou a zapamätal som si; a pomyslel som si s istým zahanbením: *Bol som hlúpy*.

*→

300. Úroveň nado mnou

Raz som požičal Xiaoguangovi „Mikeovi“ Limu moju kópiu knihy *Teória pravdepodobnosti: Logika vedy*. Mike Li si z nej niečo prečítal, potom sa vrátil a povedal:

> Fíha... je to akoby Jaynes bol tisícročný upír.

Potom Mike povedal: „Nie, počkaj, vysvetlím ako som to...“ a ja som povedal: „Nie, presne viem, čo si myslel.“ Vo fantasy literatúre platí konvencia, že čím starším je upír, tým mocnejším sa stane.

Mal som rád matematické dôkazy aj predtým, ako som sa stretol s Jaynesom. Ale E.T.Jaynes bol prvýkrát, keď som z matematických argumentov nadobudol dojem *impozantnosti*. Možno preto, lebo Jaynes menoval „paradoxy“, ktoré sa používali ako námietka na bayesiánstvo, a potom ich rozbil na kúsky ohromnou palebnou silou – silou použitou na ohromenie druhých. Alebo ten pocit impozantnosti možno pochádzal z toho, že Jaynes nebral matematiku ako estetickú hru; Jaynesovi na teórii pravdepodobnosti naozaj *záležalo*, bolo to spojené s inými vecami, na ktorých záležalo, jemu aj mne.

Z nejakého dôvodu mám z Jaynesa pocit desivo rýchlej dokonalosti – niečo, čo sa dostane k správnej odpovedi tou najkratšou možnou cestou, a pri tom pohybe zároveň všetky okolité chyby roztrhá na kúsky. Samozrejme, keď píšete knihu, máte príležitosť ukázať iba svoju najlepšiu stránku. Ale aj tak.

Hovorilo to niečo dobré aj o Mikeovi Lim, že dokázal vycítiť túto auru impozantnosti obklopujúcu Jaynesa. Ako som si všimol, že všeobecné pravidlo, že nedokážete rozlíšiť medzi úrovňami, ktoré sú príliš vysoko nad vašou vlastnou. Napríklad, raz mi niekto vážne povedal, že som naozaj bystrý, a že by som „mal ísť na vysokú“. Možno všetko, čo je viac ako jednu štandardnú odchýlku nad vami, začína splývať dokopy, aj keď toto je iba dobre znejúci divoký odhad.

Takže, keď som počul, ako Mike Li prirovnáva Jaynesa k tisícročnému upírovi, okamžite mi v hlave naskočila jedna otázka:

„Máš rovnaký pocit aj zo mňa?“ opýtal som sa.

Mike potriasol hlavou. „Prepáč,“ povedal, znejúc trochu trápne, „ale Jaynes je skrátka...“

„Nie, rozumiem,“ povedal som. Nemyslel som si, že som dosiahol Jaynesovu úroveň. Iba som bol zvedavý, ako pripadám iným ľuďom.

Usilujem sa dostať na Jaynesovu úroveň. *Usilujem* sa stať rovnakým majstrom v umelej inteligencii / reflexivite, ako je Jaynes majstrom bayesovskej teórie pravdepodobnosti. Môžem dokonca

→ http://lesswrong.com/lw/u9/my_naturalistic_awakening/

vyhlásiť, že umenie, ktoré sa chystám zvládnuť, je omnoho ťažšie než Jaynesove, čím tento rozdiel trochu nadľahčím. Ale aj tak, je to hanba, že neexistuje *žiadne* umenie, v ktorom by som bol teraz takým majstrom, ako bol Jaynes v teórii pravdepodobnosti.

To nehovorím preto, aby som sa umiestnil pod Jaynesa ako človek – aby som povedal, že Jaynes mal čarovnú auru osudu a ja nie.

Skôr v Jaynesovi vidím *úroveň odbornosti, čírej impozantnosti*, ktorú som ja ešte nedosiahol. Dokážem energicky argumentovať v mojej vybranej oblasti, ale to nie je to isté ako napísať rovnice a povedať: **HOTOVO**.

Dokiaľ som ešte nedosiahol túto úroveň, musím pripustiť možnosť, že ju nikdy nedosiahnem, že môj prirodzený talent na to nestačí. Keď ma Marcello Herreshoff poznal dosť dlho, opýtal som sa ho, či pozná niekoho, kto naňho pôsobil ako podstatne *od prírody inteligentnejší* než ja. Marcello sa na chvíľu zamyslel a povedal: „John Conway – raz som sa s ním stretol v letnom matematickom tábore.“ *Sakra*, pomyslel som si, *niekto mu napadol, a čo je horšie, je to nejaký super-slávny chlap, ktorého nemôžem len tak otravovať.* Pýtal som sa, ako Marcello k tomuto úsudku došiel. Marcello povedal: „Skrátka mi pripadalo, že má ohromné množstvo myšlienkovej konskej sily,“ a začal mi vysvetľovať nejaký matematický problém, na ktorom mal príležitosť pracovať s Conwayom.

To nebolo to, čo som chcel počuť.

Možno som vzhľadom na Marcellovu skúsenosť s Conwayon a jeho skúsenosti so mnou nemal príležitosť predviesť sa pri žiadnej téme, ktorú som zvládol rovnako dôkladne ako Conway zvládol svoje mnohé oblasti matematiky.

Alebo sa Conwayov mozog možno špecializoval iným smerom než môj, a ja sa nikdy nedokážem priblížiť ku Conwayovej úrovni v matematike, ale Conwayovi by sa zase nedarilo rovnako dobre vo výskume UI.

Alebo...

...alebo som jednoznačne hlúpejší než Conway, a on ma prevyšuje vo všetkých rozmeroch. Možno, keby som dokázal nájsť mladého proto-Conwaya a povedať mu základy, preletel by okolo mňa ako kométa, vyriešil by problémy, ktoré ma ťažia celé roky, a odfrčal by na miesta, kde by som ho nedokázal nasledovať.

Škodí môjmu egu priznať si túto poslednú možnosť? Áno. Márne by som to popieral.

Zmieril som sa *naozaj* s touto hroznou možnosťou, alebo iba sám sebe predstieram, že som sa s ňou zmieril? Tu poviem: „Nie, myslím si, že som sa s ňou zmieril.“ Čo mi dáva odvahu tak veľmi si dôverovať? Pretože som do tejto hroznej možnosti investoval konkrétne úsilie. Blogujem tu z mnohých dôvodov, ale tým hlavným je predstava, že tieto slová prečíta nejaká mladá myseľ, a prefrčí okolo mňa. Možno sa to stane, možno nie.

Alebo smutnejšie: Možno som ja vyplytval priveľa času pri zhromažďovaní prostriedkov na svoju podporu, namiesto štúdovania matematiky naplno počas celej mojej mladosti; alebo som vyplytval priveľa mladosti na nematematické myšlienky. A táto voľba, moja minulosť, sa nedá vrátiť. V štyridsiatke narazím na strop a nezostane mi nič iné, než odovzdať tieto prostriedky inej hlave s potenciálom, ktorý som ja vyplytval, stále dosť mladej na učenie sa. Aby som im ušetril čas, mal by som zanechať stopu svojich úspechov, a výstražné znamenia pri svojich chybách.

Takéto *konkrétne úsilie* založené na možnosti, ktorá zraňuje ego – to je jediný druh pokory, ktorý mi pripadá dostatočne skutočný, aby som si zaň dôveroval. Alebo vzdanie sa svojich vzácnych teórií, keď som si uvedomil, že nespĺňajú štandard, ktorý mi Jaynes ukázal – to bolo ťažké, a bolo to naozaj. Skromné správanie je lacné. Pokorné priznania pochybnosti sú lacné. Poznal som priveľa ľudí, ktorí keď počuli protiargument, povedali: „Som iba omylný smrteľník, samozrejme sa môžem mýliť,“ a potom pokračovali presne v tom istom, čo si predtým naplánovali.

Všimnete si, že sa nepokúšam skromne hovoriť nič také ako: „No, možno nie som taký skvelý ako Jaynes alebo Conway, ale to neznamená, že nedokážem robiť dôležité veci vo svojej vybranej oblasti.“

Viem totiž... že to takto nefunguje.

*→

301. Rozsah vlastnej hlúposti

V rokoch predtým než som stretol rádoby tvorcu všeobecnej umelej inteligencie (s financovaným projektom), ktorý bol náhodou kreacionista, som sa stále snažil debatovať s jednotlivými rádoby odborníkmi na VUI.

V tých časoch som v istom zmysle uspel v presviedčaní jedného takého chlapíka, že áno, musíš vziať do úvahy Priateľskú UI, a nie, nestačí iba nájsť správnu metriku spôsobilosti pre evolučný algoritmus. (Predtým naňho evolučné algoritmy urobili veľký dojem.)

A on povedal: *Ach, jaj! Ach, beda! Aký blázon som bol! Vďaka svojej neopatrnosti som takmer zničil svet! Aký zloduch som to bol!*

Nuž, *toto* je pasca, do ktorej som vedel nepadnúť...

...v tom bode, kde som sa na konci roku 2002 obzrel na návrhy UI Eliezera$_{1997}$ a uvedomil som si, čo by naozaj urobili, do tej miery, do akej boli dosť koherentné na to, aby sa dalo hovoriť o tom, čo „by naozaj urobili“.

Keď som konečne videl rozsah svojej vlastnej hlúposti, všetko hneď zapadlo na svoje miesto. Priehrada proti uvedomeniu praskla; a nevyslovené pochybnosti, ktoré sa za ňou dovtedy hromadili, prerazili všetky spolu. Nebolo tam dlhšie obdobie, dokonca si nepamätám ani jeden moment čudovania sa, ako som mohol byť taký hlúpy. Už som vedel, ako.

A tiež som vedel, všetko naraz, takpovediac v tom istom momente uvedomenia, že povedať: *takmer som zničil svet!* by bolo príliš namyslené.

Príliš by to potvrdzovalo moje ego, príliš by to potvrdzovalo moju vlastnú dôležitosť v schéme vecí, v čase, keď – pochopil som v tom istom momente uvedomenia – si moje ego zaslúžilo poriadny úder do žalúdka. Bol som omnoho menej, než som mal byť; potreboval som dostať tento úder do žalúdka, nie sa mu vyhnúť.

A rovnako som nechcel padnúť do protiľahlej pasce a povedať: *Ach, veď to nie je akoby som mal kód a chystal sa ho spustiť; nedostal som sa naozaj blízko k zničeniu sveta.* Lebo aj to by minimalizovalo silu toho úderu. *Nebolo naozaj nabité?* Navrhol som a zamýšľal som postaviť zbraň, nabiť ju, priložiť si ju k hlave a potiahnuť spúšť; a to bolo trochu veľa sebadeštruktívnosti.

Nerobil som z toho veľkú emocionálnu drámu. To by vyplytvalo veľkú silu toho úderu, premenilo ju na púhe slzy.

Vedel som, v tom istom momente, čo som opatrne ne-robil posledných šesť rokov. Neaktualizoval som.

A vedel som, že musím konečne aktualizovať. Naozaj *zmeniť* to, čo plánujem urobiť, zmeniť čo robím práve teraz, a namiesto toho robiť niečo iné.

Vedel som, že musím zastať.

Stop, roztav sa, začni horieť.

Povedz: „Nie som pripravený.“ Povedz: „Ešte neviem, ako na to.“

Toto sú hrozne ťažké slová na vyslovenie v oblasti VUI. Laické obecenstvo aj vaši spoluvýskumníci VUI sa zaujímajú o kód, projekty s pracujúcimi programátormi. Keď nemáte toto, možno vás trochu pochvália, keď poviete: „Som pripravený písať kód, len mi dajte nejaké financie.“

Povedzte: „Nie som pripravený písať kód,“ a vaše spoločenské postavenie spadne ako guľa z ochudobneného uránu.

Čo vás potom odlišuje od šiestich miliárd iných ľudí, ktorí nevedia, ako vytvoriť všeobecnú umelú inteligenciu? Ak nemáte pekný kód (ktorý samozrejme robí niečo iné než ľudsky inteligentné veci; ale aspoň je to kód), alebo minimálne svoj vlastný startup, ktorý začne písať kód hneď ako dostane financie – kto potom ste, a čo robíte na našej konferencii?

→ http://lesswrong.com/lw/ua/the_level_above_mine/

Možno neskôr napíšem článok o tom, odkiaľ pochádza tento postoj – chýbajúci stred medzi: „Viem, ako postaviť VUI!“ a „Pracujem na užšej UI, pretože neviem, ako postaviť VUI“, neexistencia pojmu pre: „Snažím sa dostať od neúplnej mapy Priateľskej UI k úplnej mape Priateľskej UI“.

Lenže tento postoj existuje, a tak je strata spoločenského postavenia spojená so slovami: „Nie som pripravený písať kód“ veľmi veľká. (Ak o tom niekto pochybuje, menujte mi niekoho iného, kto zároveň hovorí: „Chcem zostrojiť všeobecnú umelú inteligenciu“, „Práve teraz nemôžem zostrojiť VUI, pretože neviem X“ a „Práce pracujem na skúmaní X“.)

(A to nehovorím o ľuďoch z VUI, ktorí už získali rizikový kapitál a sľubujú návratnosť o päť rokov.)

Takže existuje veľká neochota povedať „Stop“. Nemôžete jednoducho povedať: „Ach, vrátim sa späť do režimu skúmania X“, pretože taký režim neexistuje.

Bolo v mojom prípade za touto neochotou niečo viac než len strata spoločenského postavenia? Eliezer_{2001} mohol aj cúvnuť pred spomalením svojej vnímanej zotrvačnosti smerom k explózii inteligencie, ktorá bola taká správna a taká potrebná...

Ale myslím si, že som najmä cúvol pred nemožnosťou povedať: „Som pripravený začať kódovať.“ Nie iba zo strachu z reakcie druhých, ale preto, že som mal sám vštepený rovnaký postoj.

Najmä, Eliezer_{2001} nepovedal „Stop“ - dokonca ani *po* všimnutí si problému s Priateľskou UI – pretože som si neuvedomil, na úrovni inštinktu, že Príroda má dovolené zabiť ma.

„Puberťáci si myslia, že sú nesmrteľní,“ hovorí príslovie. Samozrejme to nie je pravda v doslovnom zmysle, že ak sa ich opýtate: „Si nezničiteľný?“ odpovedia vám: „Áno, poď a skús ma zastreliť.“ Ale možno pre nich pripásanie sa v aute nie je emocionálne dôležité, pretože myšlienka na vlastnú smrť nie je celkom *skutočná* – neveria, že je naozaj dovolené, aby sa to stalo. Môže sa to stať *v princípe*, ale nemôže sa to stať *naozaj*.

Ja som osobne vždy bol pripásaný. Ako jednotlivec som chápal, že môžem zomrieť.

Ale keďže som bol vychovaný v technofílii a cenil som si tú najvzácnejšiu vec, omnoho dôležitejšiu než svoj vlastný život, myslel som si kedysi, že Budúcnosť je nezničiteľná.

Dokonca aj keď som uznal, že nanotechnológia by mohla zničiť ľudstvo, stále som veril, že explózia inteligencie bude nezraniteľná. Že keby ľudstvo prežilo, nastala by explózia inteligencie, a bola by príliš bystrá na to, aby sme ju pokazili alebo stratili.

Ešte aj *po tom*, čo som uznal Priateľskú UI ako úvahu, emocionálne som neveril v možnosť zlyhania, rovnako ako puberťák, ktorý si nezapína bezpečnostný pás, neverí *naozaj*, že automobilová nehoda má *naozaj* dovolené ho zabiť alebo zmrzačiť.

Dokiaľ mi môj vhľad do optimalizácie neumožnil obzrieť sa a vidieť Eliezera_{1997} v jasnom svetle, neuvedomil som si, že Príroda má dovolené zabiť ma.

„Myšlienka, ktorú si nemôžeš pomyslieť, ťa ovláda viac než myšlienky, ktoré hovoríš nahlas.“ Ale my cúvame iba pred tými obavami, ktoré sú pre nás skutočné.

Výskumníci VUI berú veľmi vážne možnosť, že *niekto iný vyrieši tento problém ako prvý*. Dokážu si predstaviť, ako vidia novinové titulky hovoriace, že ich vlastné dielo sa ocitlo v pozadí. Vedia, že Príroda má dovolená urobiť im toto. Tí, ktorí si založili firmy vedia, že je dovolené, aby sa im minul rizikový kapitál. Táto možnosť je pre nich *skutočná*, celkom skutočná; má nad nimi moc emocionálneho nutkania.

Nemyslím si, že „Joj“, po ktorom by nasledoval buchot šiestich miliárd padajúcich tiel, *ich vlastnou rukou*, je pre nich skutočné na celkom rovnakej úrovni.

Nie je bezpečné hovoriť, čo si myslia iní ľudia. Ale zdá sa celkom pravdepodobné, že keď niekto reaguje na predstavu Priateľskej UI slovami: „Ak spomalíš vývoj, aby si pracoval na bezpečnosti, iné projekty, ktorým *vôbec* nezáleží na Priateľskej UI ťa predbehnú,“ predstava, že osobne urobia chybu, po ktorej bude nasledovať šesť miliárd buchnutí pre nich nie je celkom skutočná; ale možnosť, že by ich niekto predbehol, je hlboko desivá.

Aj ja som kedysi hovoril takéto veci, než som pochopil, že Príroda má dovolené zabiť ma.

V tom momente uvedomia sa moja technofília z detstva konečne zlomila.

Konečne som pochopil, že ešte aj keď usilovne dodržiavate pravidlá vedy a ste dobrý človek, Príroda vás stále môže zabiť. Konečne som pochopil, že aj keby ste boli ten najlepší projekt zo všetkých možných kandidátov, Príroda vás stále môže zabiť.

Pochopil som, že ma neznámkujú podľa rebríčka. Môj pohľad sa zbavil súperov, a uvidel som čistý prázdny múr.

Obzrel som sa a videl som tie opatrné argumenty, ktoré som zostavil, prečo je najmúdrejšou možnosťou pokračovať plnou parou vpred, rovnako ako som plánoval predtým. A pochopil som, že dokonca aj keď zostavíte argument ukazujúci, že niečo je najlepší smer, Príroda má stále dovolené povedať „No a čo?“ a zabiť vás.

Obzrel som sa a videl som, že som tvrdil, že beriem do úvahy riziko principiálnej chyby, že som tvrdil, že sú dôvody prečo tolerovať riziko postupovania v neprítomnosti plného poznania.

A videl som, že to riziko, ktoré som chcel tolerovať, by ma bolo zabilo. A videl som, že mi táto možnosť nikdy nepripadala *skutočná*. A videl som, že dokonca aj keby ste mali múdre a vynikajúce argumenty pre riskovanie, toto riziko malo stále dovolené ísť a zabiť vás. *Naozaj* vás zabiť.

Pretože záleží iba na činoch, a nie na dôvodoch, prečo niečo robíte. Ak postavíte zbraň, nabijete zbraň, priložíte si zbraň k hlave a stlačíte spúšť, aj keď máte tie najchytrejšie argumenty pre vykonanie každého kroku – potom, bum.

Videl som, že iba moja vlastná neznalosť pravidiel mi umožnila hádať sa, že treba ísť vpred bez úplnej znalosti pravidiel; pretože ak nepoznáte tie pravidlá, nedokážete si predstaviť pokutu za nevedomosť.

Videl som, ako ostatní, stále nevedomí pravidiel, hovorili: „Pôjdem vpred a urobím X“; a že do tej miery, do akej bolo X vôbec koherentným návrhom, vedel som, že výsledkom by bol tresk; ale oni povedali: „Neviem, či by to nemohlo fungovať“. Skúšal by som im vysvetliť, aký malý je terč vo vyhľadávacom priestore, a oni by povedali: „Ako si môžeš byť taký istý, že nevyhrám v lotérii?“, držiac svoju vlastnú nevedomosť ako obušok.

A tak som si uvedomil, že jediná vec, ktoré by som *mohol* urobiť, aby som sa zachránil vo svojom predchádzajúcom stave nevedomosti, bola povedať: „Nebudem pokračovať, dokiaľ nebudem pozitívne vedieť, že je to bezpečné územie.“ A existuje veľa chytrých argumentov, prečo by ste mali stúpiť na územie, o ktorom neviete, či obsahuje nášľapnú mínu; ale všetky znejú omnoho menej chytro po tom, čo sa pozriete na miesto, na ktoré ste odporúčali a zamýšľali stúpiť, a vidíte ten tresk.

Pochopil som, že môžete urobiť *všetko, čo ste mali urobiť*, a Príroda má aj tak dovolené vás zabiť. Tak sa zlomila moja posledná dôvera. A vtedy sa začal môj výcvik ako racionalistu.

* →

—

302. Mimo dosahu Boha

Dnešný článok je o čosi pochmúrnejší než zvyčajne, podľa mojej mierky. Týka sa myšlienkového experimentu, ktorý som vymyslel, aby som rozbil svoj vlasrný optimizmus, keď som si uvedomil, že ma optimizmus zviedol z cesty. Čitatelia, ktorí prikyvujú argumentom ako: „Je dôležité ponechať si svoje skreslenia, pretože nám pomáhajú zostať šťastnými,“ by mali radšej prestať čítať. (Pokiaľ nemajú niečo, na čom im záleží, vrátane svojho vlastného života.)

Takže! Hľadiac späť na rozsah svojej vlastnej hlúposti, uvedomil som si, že jej koreňom bola neviera v zraniteľnosť Budúcnosti – neochota prijať, že veci *naozaj* môžu dopadnúť zle. Nie ako dôsledok nejakého explicitného výrokového názoru. Skôr ako niečo vnútri, čo vytrvalo vo viere, že napriek nepriazni osudu môže bude nakoniec všetko v poriadku.

Niektorí by toto považovali za cnosť (*zettai daijobu da yo*) a iní by povedali, že je to nevyhnutné pre duševné zdravie.

→ http://lesswrong.com/lw/ue/the_magnitude_of_his_own_folly/

Lenže my nežijeme v takom svete. Žijeme vo svete mimo dosahu Boha.

Bolo to dávno, veľmi dávno, keď som veril v Boha. Vyrástol som v ortodoxnej židovskej rodine, a viem si spomenúť na posledný zapamätaný raz, keď som Boha o niečo žiadal, hoci si nepamätám, koľko som mal rokov. Prihováral som sa v niečom za susedovie chlapca, už som zabudol, čo presne – niečo v duchu „dúfam, že dopadne dobre“ alebo možno „dúfam, že sa stane židom“.

Pamätám sa, aké to bolo mať nejakú vyššiu autoritu, ku ktorej sa dalo odvolať, aby sa postarala o veci, ktoré som nedokázal sám zvládnuť. Nemyslel som na to, ako na „hrejivý“ pocit, pretože som nemal žiadnu alternatívu, s ktorou by som to porovnal. Jednoducho som to bral ako samozrejmosť.

Každopádne si spomínam, hoci iba zo vzdialeného detstva, aké to bolo žiť v pojmovo nemožnom možnom svete, kde existoval Boh. *Naozaj* existoval, v tom zmysle ako deti a racionalisti berú všetky svoje názory doslovne.

Vo svete, kde Boh existuje, zasahuje, aby optimalizoval *všetko*? Bez ohľadu na to, čo rabíni tvrdia o základnej podstate skutočnosti, seriózna operatívna odpoveď na túto otázku je samozrejmé: „Nie.“ Nežiadate Boha, aby vám priniesol limonádu z chladničky, namiesto aby ste si ju vzali sami. Keď som veril v Boha vážnym detským spôsobom, tak veľmi dávno, neveril som v toto.

Keď predpokladáme toto, konkrétna božská neaktivita nevyvoláva úplnú teologickú krízu. Keby ste mi povedali: „Zostrojil som dobrotivého superinteligentného používateľa nanotechnológie“ a ja by som povedal: „Daj mi banán“ a žiaden banán by sa neobjavil, to by *ešte* nevyvracalo vaše tvrdenie. Ľudskí rodičia nerobia vždy to, o čo ich deti požiadajú. V teórii zábavy sú nejaké slušné argumenty – dokonca im sám verím – proti predstave, že *najlepšia* pomoc, akú niekomu môžete poskytnúť, je vždy mu ihneď dať všetko, čo chce. Nemyslím si, že eudaimonia je formulovať ciele a mať ich okamžite splnené; *nechcel* by som sa stať jednoduchou želajúcou si vecou, ktorá nikdy nepotrebuje plánovať alebo konať alebo myslieť.

Nie je teda nevyhnutne pokusom vyhnúť sa falzifikácii, ak povieme, že Boh neplní všetky modlitby. Dokonca ani Priateľská UI by neodpovedala na každú požiadavku.

Ale iste existuje *nejaká* hranica hrôzy dostatočne strašnej na to, aby Boh zasiahol. Toto si pamätám ako pravdu, keď som veril detským spôsobom.

Boh, ktorý nezasiahne *nikdy*, bez ohľadu na to, čo zlé sa stane – *to* je jasný pokus vyhnúť sa falzifikácii, chrániť vieru vo vieru. Dostatočne malé dieťa nemá poznanie hlboko vnútri, že Boh naozaj neexistuje. Naozaj očakáva, že vo svojej garáži uvidí draka. Nemá dôvod predstavovať si milujúceho Boha, ktorý nikdy nekoná. Kde presne je tá hranica dostatočnej strašnosti? Dokonca aj dieťa si vie predstaviť hádku ohľadom presnej hranice. Ale Boh samozrejme tú čiaru niekde urobí. Je predsa málo milujúcich rodičov, ktorí by z túžby, aby ich dieťa vyrástlo silné a samostatné, nechali svoje batoľa zraziť autom.

Samozejmým príkladom hrôzy takej veľkej, že by ju Boh nemohol dovoliť, je smrť – skutočná smrť, zničenie mysle. Nemyslím, že to dovoľuje dokonca ani buddhizmus. Dokiaľ teda existuje Boh v klasickom zmysle – plnohodnotný, ontologicky základný, *ten* Boh – môžeme byť pokojní, že sa žiadna *dostatočne* strašná udalosť nikdy, nikdy nestane. Neexistuje duša, ktorá by sa niekde musela báť skutočného zničenia; Boh by tomu zabránil.

Čo ak si postavíte svoj vlastný simulovaný vesmír? Klasickým príkladom simulovaného vesmíru je Conwayova hra Život. Vyzývam vás, aby ste preskúmali Život, ak ste ho nikdy nehrali – je to dôležité na pochopenie pojmu „fyzikálny zákon“. Conwayov život je dokázateľne Turingovsky úplný, takže by bolo možné vo vesmíre Života zostaviť vedomú bytosť, hoci by to mohlo byť veľmi krehké a čudné. S inými bunkovými automatmi by to bolo ľahšie.

Mohli by ste pomocou vytvorenia simulovaného vesmíru uniknúť z dosahu Boha? Mohli by ste simulovať hru Život obsahujúcu vedomé bytosti, a mučiť tieto bytosti? Keby však Boh všetko sledoval, potom by pokus o vybudovanie neférového Života akurát spôsobil, že by *ten* Boh zasiahol, aby upravil tranzistory na vašom počítači. Ak fyzika, ktorú nastavíte vo svojom počítačovom programe, žiada, aby bola vedomá bytosť v Živote donekonečna bezdôvodne mučená, potom *ten* Boh zasiahne. Boh je všadeprítomný, neexistuje *žiadne* útočisko pre skutočnú hrôzu: Život je férový.

Ale predstavte si, že sa vás namiesto toho opýtam:

Ak sú dané také a také počiatočné podmienky, a *ak sú dané* také a také pravidlá bunkového automatu, čo *by bolo* matematickým výsledkom?

Ani Boh nemôže zmeniť odpoveď na túto otázku, pokiaľ teda neveríte, že Boh môže uskutočniť logické nemožnosti. Nepamätám sa, že by som v toto veril ani ako veľmi malé dieťa. (A aký dôvod by ste mali v to veriť, keď Boh môže zmeniť všetko, čo *naozaj* existuje?)

Ako vyzerá Život, v tomto imaginárnom svete, kde každý krok vyplýva *iba* zo svojho bezprostredného predchodcu? Kde sa veci iba dejú, alebo nedejú, *iba* kvôli pravidlám bunkového automatu? Kde počiatočné podmienky a pravidla *neopisujú* žiadneho Boha, ktorý kontroluje každý stav? Ako to vyzerá, v tom svete mimo dosahu Boha?

Ten svet by nebol férový. Keby počiatočný stav obsahoval zárodok niečoho, čo by sa mohlo replikovať, prirodzený výber by mohol alebo nemohol nastať, a zložitý život by sa mohol alebo nemohol vyvinúť, a ten život by sa mohol alebo nemohol stať vedomým, pričom by žiaden Boh tento vývoj neriadil. Ten svet by mohol vyvinúť ekvivalent vedomých kráv, alebo vedomých delfínov, ktorým by chýbali ruky, aby mohli zlepšiť svoju situáciu; možno by ich jedli vedomé vlky, ktoré by si nikdy nepomysleli, že robia niečo zlé, alebo by im na tom nezáležalo.

Keby sa v rozľahlej palete možností vyvinulo niečo ako ľudia, potom by trpeli chorobami – nie preto, aby ich to naučilo nejakú lekciu, ale iba pretože sa podľa pravidiel bunkového automatu vyvinuli aj vírusy.

Ak sú ľudia v tom svete šťastní alebo nešťastní, príčiny ich šťastia alebo nešťastia nemusia mať nič spoločné s dobrými alebo zlými voľbami, ktoré urobili. Nič spoločné so slobodnou vôľou, alebo naučenými lekciami. V tomto imaginárnom svete, kde každý krok vyplýva iba z pravidiel bunkového automatu, ekvivalent Džingischána môže zavraždiť milión ľudí, a smiať sa, byť bohatý, a nikdy nebyť potrestaný, a žiť svoj život omnoho šťastnejšie než priemer. Kto tomu zabráni? Boh by samozrejme zabránil, aby sa to niekedy *naozaj* stalo; prinajmenšom by do chánovho srdca poslal nejaký pocit stiesnenosti. Ale v matematickej odpovedi na otázku *Čo keby?* neexistuje žiaden Boh v axiómach. Ak teda pravidlá bunkového automatu povedia, že chán je šťastný, to je jednoducho celá a jediná odpoveď na otázku čo keby. Neexistuje nič, absolútne nič, čo by tomu zabránilo.

Ak čo chán mučí ľudí hrozným spôsob na smrť, celé dni, napríklad pre svoje vlastné pobavenie? Budú volať o pomoc, možno si budú predstavovať Boha. A keby ste *naozaj* napísali taký bunkový automat, Boh by samozrejme mohol vo vašom programe zasiahnuť. Ale v otázke čo keby, čo *by* ten bunkový automat urobil za daných matematických pravidiel, tam nie je v systéme žiaden Boh. Keďže fyzikálne zákony neobsahujú žiadnu špecifikáciu funkcie úžitku – konkrétne, žiaden zákaz mučenia – potom možno obete zachrániť iba ak tie správne bunky budú mať hodnotu 0 alebo 1. A je nepravdepodobné, že sa niekto postaví chánovi na odpor; keby áno, niekto by ho sekol mečom, meč by mu roztrhol orgány, a on by zomrel, a to by bol koniec. Takže obete zomierajú, kričiac, a nikto im nepomáha; to je odpoveď na otázku čo keby.

Mohli by tie obete byť celkom nevinné? Prečo nie, vo svete čo keby? Ak sa pozriete na pravidlá Conwayovej hry Život (ktorá je Turingovsky úplná, takže do nej môžeme vložiť ľubovoľnú vypočítateľnú fyziku), potom sú pravidlá naozaj veľmi jednoduché. Bunky s troma zapnutými susedmi budú zapnuté; bunky s dvoma zapnutými susedmi zostanú rovnaké, všetky ostatné bunky sa vypnú. Nie je tam nič o tom, že nevinných ľudí nemožno neobmedzene dlho strašne mučiť.

Začína vám tento svet znieť povedome?

Viera v spravodlivý vesmír sa často prejavuje jemnejším spôsobon, než myslením si, že hrôzy by mali byť priamo zakázané: Odohralo by sa dvadsiate storočie inak, keby sa Klara Pölzl a Alois Hitler boli milovali o hodinu skôr, a iná spermia by bola oplodnila vajíčko, v tú noc, keď bol počatý Adolf Hitler?

Aby tak veľa životov a toľko straty záviselo od jedinej udalosti, vyzerá *neprimerane*. Božský plán by mal dávať viac *zmyslu* než toto. Môžete veriť v Božský plán a pritom neveriť v Boha – Karl Marx to tak iste robil. Nemalo by to byť tak, že milióny životov závisia od náhodnej voľby, od načasovania, od rýchlosti mikroskopického bičíka. Nemalo by to byť dovolené. Je to príliš *neprimerané*. Preto, keby

Adolf Hitler mohol ísť na strednú školu a stať sa architektom, jeho rolu by bol prevzal niekto iný, a druhá svetová vojna by sa bola odohrala rovnako ako predtým.

Lenže vo svete mimo dosahu Boha, neexistuje vo fyzikálnych axiómach žiadna klauzula, ktorá hovorí „veci musia dávať zmysel“ alebo „veľké účinky si vyžadujú veľké príčiny“ alebo „dejiny sa riadia dôvodmi príliš dôležitými na to, aby boli také krehké“. Nie je tam žiaden Boh, ktorý by si *vynútil* tento poriadok, ktorý je tak vážne porušený, keď životy a smrti miliónov závisia od jednej malej molekulárnej udalosti.

Pointa tohto myšlienkového experimentu je položiť vesmír s Bohom a vesmír s Prírodou vedľa seba, aby sme dokázali rozoznať, ktorý druh myslenia patrí do vesmíru s Bohom. Mnohí ateisti stále rozmýšľajú, akoby isté veci *neboli dovolené*. Zostavujú argumenty, prečo bola druhá svetová vojna nevyhnutná, a bola by sa odohrala viacmenej rovnako aj keby sa Hitler bol stal architektom. Ale v triezvych historických faktoch je toto nerozumný názor; vybral som si druhú svetovú vojnu ako príklad, pretože podľa toho, čo som čítal, to vyzerá, že jej udalosti boli prevažne určované Hitlerovou osobnosťou, často vzdorujúcou svojim generálom a poradcom. Neexistuje žiadne konkrétne empirické zdôvodnenie, o ktorom by som počul, prečo o tomto pochybovať. Hlavným dôvodom pochybovať je *odmietnutie prijať*, že vesmír môže dávať tak málo zmyslu – že sa hrozné veci môžu stať tak *ľahko*, nemajúc viac dôvodu než hodenie kockou.

Ale prečo nie? Čo to zakazuje?

Vo vesmíre s Bohom to zakazuje Boh. Uvedomiť si toto znamená uvedomiť si, že nežijeme v takom vesmíre. Žijeme vo vesmíre čo keby, mimo dosahu Boha, riadenom matematickými zákonmi a ničím iným. O čom fyzika povie, že sa to stane, to sa stane. Absolútne *hocičo*, dobré alebo zlé, sa stane. A neexistuje nič vo fyzikálnych zákonoch, čo by urobilo výnimku z tohto pravidla aspoň pre tie *naozaj extrémne* prípady, kde by ste mohli očakávať, že sa Príroda zachová trochu rozumnejšie.

Keď si čítam *Vzostup a pád tretej ríše* od Williama Shirera, počúvam ako opisuje nedôveru, ktorú on a ďalší cítili, keď objavili celý rozsah nacistických zverstiev, pomyslel som si, aké je to zvláštne, čítať to všetko a už vedieť, že proti tomu neexistuje jediná ochrana. Jednoducho prečítať celú túto knihu a prijať to; zdesený, ale vôbec nie neveriaci, pretože už som chápal, v akom druhu sveta žijem.

Kde bolo, tam bolo, kedysi som veril, že vyhynutie ľudstva nie je dovolené. A ostatní, ktorí si hovoria racionalisti, možno tiež ešte veria v nejaké veci. Môžu ich nazývať „hry s nulovým súčtom“ alebo „demokracia“ alebo „technológia“, ale sú posvätné. Známkou tejto posvätnosti je, že táto dôveryhodná vec nemôže viesť k ničomu *naozaj* zlému; alebo nemôže byť *trvalo* skazená, prinajmenšom nie bez zodpovedajúceho svetla na konci tunela. V tom zmysle jej možno dôverovať, aj keď sa tu a tam stane pár zlých vecí.

Odvíjajúce sa dejiny Zeme sa nemôžu nikdy obrátiť od svojho trendu pozitívnych súčtov k trendu negatívnych súčtov; to nie je dovolené. Demokracie – prinajmenšom *moderné liberálne* demokracie – by nikdy nelegalizovali mučenie. Technológia urobila doteraz toľko dobrého, že nie je možné, aby existovala nejaké technológia čiernej labute, ktorá poruší tento trend a urobí viac škody než bolo všetkého dobrého až doteraz.

Existujú všelijaké chytré argumenty, prečo sa takéto veci nemôžu ani náhodou stať. Ale zdrojom týchto argumentov je omnoho hlbšia viera, že takéto veci *nie sú dovolené*. Ale kto ich zakazuje? Kto bráni, aby nastali? Ak si nedokážete predstaviť aspoň jeden zákonitý vesmír, v ktorom fyzika hovorí, že takéto hrozné veci sa stávajú – a teda *sa stávajú*, lebo sa niet kam odvolať voči rozsudku – potom nie ste naozaj pripravení diskutovať o *pravdepodobnostiach*.

Mohlo by to byť naozaj tak, že vedomé bytosti definitívne zomierali tisíce alebo milióny rokov, bez žiadnej duše ani posmrtného života – *ani* ako súčasť žiadneho veľkého plánu Prírody – *nie* aby sa naučili nejakú veľkú lekciu o zmysluplnosti alebo nezmyselnosti života – ani by sa naučili nejakú hlbokú lekciu o tom, čo je nemožné – a že trik taký jednoduchý a hlúpo znejúci ako zmrazenie ľudí v tekutom dusíku ich môže zachrániť pred úplným zničením – a 10-sekundové odmietnutie tejto hlúpej myšlienky môže zničiť niekoho dušu? Môže to byť tak, že programátor, ktorý podpíše pár tlačív a kúpi si životné poistenie pokračuje do vzdialenej budúcnosti, zatiaľ čo Einstein hnije v hrobe? Jednou vecou si môžeme byť istí:

Boh by to nedovolil. Čokoľvek také smiešne a neprimerané by iste bolo vylúčené. Bol by to výsmech Božského plánu – výsmech *silných dôvodov*, prečo veci musia byť také, aké sú.

Vy možno máte svetské racionalizácie, prečo veci *nie sú dovolené*. Možno teda pomôže predstaviť si, že existuje nejaký Boh, dobrý v to zmysle, ako vy rozumiete dobro – Boh, ktorý v celej Skutočnosti vynucuje aspoň *minimum* spravodlivosti – ktorého plány dávajú zmysel a primerane závisia od voľby ľudí – ktorý by nikdy nedovolil absolútnu hrôzu – ktorý nezasahuje vždy, ale ktorý prinajmenšom nedovolí vesmíru, aby sa *celkom* vykoľajil zo svojej dráhy... predstavte si toto všetko, ale predstavte si aj, že *vy* osobne žijete vo svete čo keby z čírej matematiky – vo svete mimo dosahu Boha, v úplne nechránenom svete, kde sa môže stať absolútne čokoľvek.

Ak toto ešte číta niekto, kto si myslí, že byť šťastný znamená viac než hocičo v živote, možno by potom *nemal* tráviť veľa času rozjímaním o nechránenosti svojej existencie. Možno by mal na ňu myslieť *len* dosť dlho na to, aby zapísal seba a svoju rodinu na kryoniku, a/alebo občas napísal šek nejakej agentúre na zmierňovanie existenciálneho rizika. A mal by byť v aute pripásaný, a mať zdravotné poistenie, a všetky tie ďalšie smutné nevyhnutnosti, ktoré dokážu zničiť váš život, ak vynecháte jeden krok... ale okrem toho, ak chcete byť šťastní, meditovanie o krehkosti života vám nepomôže.

Tento článok bol však napísaný pre tých, ktorí majú čo chrániť.

Čo môže urobiť roľník v dvanástom storočí, aby sa zachránil pred zničením? Nič. Problémy predložené Prírodou nie sú vždy fér. Keď natrafíte na problém, ktorý je príliš zložitý, utrpíte trest; keď narazíte na smrteľný trest, zomriete. Tak to funguje pre ľudí, a funguje to rovnako aj pre planéty. Kto chce tancovať smrtiaci tanec Prírody, musí rozumieť, voči čomu tu stojí: Absolútnej, naprostej neutralite bez výnimiek.

Vedieť toto vás vždy nezachráni. Nezachránilo by to roľníka v dvanástom storočí, ani keby to vedel. Ak si mylíte, že racionalista, ktorý celkom chápe, v akej kaši sa nachádza, musí *určite* dokázať nájsť cestu von – potom dôverujete rozumnosti, škoda komentovať.

Niektorý diskutér ma iste bude hrešiť za to, že tomu celému dávam príliš temný tón, a ako reakciu vymenuje všetky dôvody, prečo je krásne žiť v neutrálnom vesmíre. Život má predsa povolené byť *trochu* temný; ale nie temnejší za istú hranicu, pokiaľ tam nie je svetlo na konci tunela.

Napriek tomu, keďže nechcem vytvárať *zbytočné* zúfalstvo, poviem k tomu pár slov nádeje:

Ak sa budúcnosť ľudstva odvinie tým správnym smerom, možno sa nám podarí urobiť náš budúci svetelný kužeľ spravodlivým (spravodlivejším). Nemôžeme zmeniť základnú fyziku, ale na vyšších úrovniach organizácie môžeme postaviť nejaké zábradlia a dať naspodok nejaké vypchávky; zorganizovať častice do vzorov, ktoré majú nejakú vnútornú kontrolu proti katastrofe. Existuje veľa vecí, na ktoré nedosiahneme – ale možno pomôže, ak si predstavíme všetko, čo nie je v našom budúcom svetelnom kuželi ako časť „zovšeobecnenej minulosti“. Akoby sa to všetko už stalo. Prinajmenšom je tu *vyhliadka*, že porazíme neutralitu v tej jedinej budúcnosti, na ktorú dosiahneme – v jedinom svete, kde môžeme niečo dosiahnuť tým, že sa o to budeme starať.

Jedného dňa budú možno nezrelé mysle spoľahlivo chránené. Hoci deti zažijú ekvivalent nedostania lízatka alebo možno aj popálenia si prstu, nikdy ich nezrazí auto.

A dospelí nebudú v toľkom nebezpečenstve. Superinteligenciu – myseľ, ktorá by dokázala myslieť bilión myšlienok bez jediného chybného kroku – by nezastrašil problém, kde sa za jediné zlyhanie platí smrťou. Surový vesmír by nevyzeral taký drsný, bol by to iba ďalší problém na vyriešenie.

Problém je, že zostrojiť dospelého je samo osebe problémom pre dospelých. Toto som si pred rokmi konečne uvedomil.

Ak existuje spravodlivý (spravodlivejší) vesmír, musíme sa tam dostať začínajúc z *tohto* sveta – neutrálneho sveta, sveta s tvrdým betónom bez vypchávok, sveta, kde problémy nie sú kalibrované podľa vašich schopností.

Nie každá dieťa musí pozerať Prírode do očí. Zapnúť si pás, alebo napísať šek, nie je také zložité ani smrtiace. Nehovorím, že každý racionalista by mal meditovať o neutralite. Nehovorím, že každý racionalista by mal myslieť na všetky tieto nepríjemné myšlienky. Ale nikto, kto plánuje čeliť nekalibrovanej hrozbe okamžitej smrti, sa im nesmie vyhýbať.

Čo musí urobiť dieťa – aké pravidlá by malo dodržiavať, ako by sa malo správať – aby vyriešilo problém pre dospelých?

*→

303. Moje bayesovské osvietenie

Spomínam si (hmlisto, ako to pri ľudských spomienkach býva), keď som sa po prvýkrát sám identifikoval ako „bayesovec“. Niekto sa práve opýtal nesprávne formulovanú verziu starého hlavolamu o pravdepodobnostiach, slovami:

> Ak stretnem na ulici matematičku a ona mi povie: „Mám dve deti, a aspoň jedno z nich je chlapec,“ aká je pravdepodobnosť, že sú obaja chlapci?

V *správnej* verzii tohto príbehu matematička povie: „Mám dve deti,“ a *vy* sa opýtate: „Je aspoň jedno z nich chlapec?“ a ona odpovie: „Áno“. Potom je pravdepodobnosť 1/3, že sú obaja chlapci.

Ale v tejto nesprávne formulovanej verzii príbehu – ako som poukázal – by človek prirodzene uvažoval takto:

> Keby tá matematička mala jedného chlapca a jedno dievča, potom moja pôvodná pravdepodobnosť, že povie „aspoň jedno z nich je chlapec“ je ½, a môja pôvodná pravdepodobnosť, že povie „aspoň jedno z nich je dievča“ je ½. Nemám dôvod veriť, a priori, že tá matematička spomenie dievča iba vtedy, ak neexistuje iná možnosť.

Takže som poukázal na toto a vypočítal odpoveď podľa Bayesovho pravidla, dostal som pravdepodobnosť ½, že obe deti sú chlapci. Nie som istý, či už som vtedy vedel alebo nie, že sa to volá Bayesovo pravidlo, ale použil som ho.

A hľa, niekto mi povedal: „No, ty si práve dal bayesovskú odpoveď, ale podľa klasickej štatistiky je odpoveď 1/3. Iba vylúčime možnosti, ktoré sú v rozpore s danými informáciami, a spočítame tie, ktoré zostali, a nesnažíme sa uhádnuť pravdepodobnosť, že tá matematička povie to alebo ono, pretože tú pravdepodobnosť nemáme v skutočnosti ako zistiť – je príliš subjektívna.“

Odpovedal som – podotýkam, že celkom spontánne - „Čo tým akože myslíš? Priradeniu pravdepodobnosti tomu, že tá matematička povie jednu vetu a nie druhú, sa predsa nemôžeš vyhnúť. Ty akurát predpokladáš, že tá pravdepodobnosť je 1, a *na to* nemáš dôvod.“

Na čo dotyčný odpovedal: „Áno, to hovoria bayesovci. Ale frekventisti si to nemyslia.“

A ja som povedal, prekvapene: „Ako vôbec môže existovať niečo také ako ne-bayesovská štatistika?“

Vtedy som objavil, že patrím k tomu typu, ktorý sa nazýva „Bayesovci“. Pokiaľ vie, už som sa takto *narodil*. Moja matematická intuícia bola vždy taká, že všetko, čo povedali Bayesovci, mi pripadalo dokonale priamočiare a jednoduché, samozrejmý spôsob, akým by som to robil ja sám; zatiaľ čo veci, ktoré hovorili frekventisti, zneli ako komplikované, pokrivené, šialené rúhanie snívajúceho Cthulhu. *Nevybral* som si, že bude Bayesovcom o nič viac, než si ryby vyberajú, že budú dýchať vo vode.

Ale toto nie je to, čo označujem ako svoje „Bayesovské osvietenie“. Prvýkrát, keď som počul o „Bayesiánstve“, som si to odškrtol ako samozrejmosť; nešiel som omnoho ďalej než po samotné Bayesovo pravidlo. V tom čase som stále myslel na teóriu pravdepodobnosti ako na nástroj, nie zákon. Nemyslel som si, že existujú matematické zákony pre inteligenciu (moja najlepšia a najhoršia chyba). Podobne ako takmer všetci rádoby odborníci na VUI, Eliezer_{2001} rozmýšľal v pojmoch techník, metód, algoritmov, budovania sady nástrojov plnej cool vecí, ktoré by mohol *urobiť*; hľadal nástroje, nie porozumenie. Bayesovo pravidlo bolo naozaj šikovný nástroj, použiteľný v prekvapivo mnohých prípadoch.

Potom som bol zasvätený do heuristík a skreslení. Začalo to, keď som natrafil na webovú stránku, ktorá bola konvertovaným powerpointovým úvodom do behaviorálnej ekonomiky. Spomínala nejaké

→ http://lesswrong.com/lw/uk/beyond_the_reach_of_god/

výsledky heuristík a skreslení, len tak mimochodom, bez odkazov. Bol som taký ohromený, že som poslal autorovi e-mail s otázkou, či to bol skutočný experiment alebo iba anekdota. Poslal mi naspäť sken článku Tverskeho a Kahnemana z roku 1973.

Hanbím sa, že musím povedať, že tam môj príbeh naozaj nezačal. Odložil som si to do svojho zoznamu vecí, na ktoré sa pozriem neskôr. Vedel som, že existuje zbierka „Úsudok pri neistote: Heuristiky a skreslenia“, ale nikdy som ju nevidel. V tom čase som došiel k záveru, že keď to nie je na internete, tak si jednoducho skúsim poradiť bez toho. Mal som vo svojom zozname na čítanie veľa iných vecí, a nemal som ľahký prístup do univerzitnej knižnice. Asi som to spomenul v mailovej skupine, pretože Emil Gilliam bol taký naštvaný z mojej teórie o čítaní iba internetu, že mi tú knihu kúpil.

Tento jeho čin by si pravdepodobne zaslúžil slušný počet bodov.

Ale ani toto nie je to, čo označujem za svoje „Bayesovské osvietenie“. Bol to dôležitý krok smerom k uvedomeniu si neprimeranosti svojich zručností Tradičnej Rozumnosti – že toho existuje oveľa viac, celá táto nová veda, za tým, čo vám Richard Feynman povedal, že máte robiť. A vidieť, že program heuristík a skreslení považuje Bayesa za zlatý štandard pomohlo posunúť moje myslenie vpred – ale nie až celkom tam.

Pamäť je krehká vec, a moja sa zdá byť ešte krehkejšia než väčšina, odkedy som sa naučil, ako sa spomienky znovuvytvárajú pri každom novom spomínaní – veda o tom, aké hrehké sú naozaj. Majú iní ľudia naozaj lepšiu pamäť, alebo jednoducho dôverujú podrobnostiam, ktoré si ich pamäť vymyslí, zatiaľ čo si v skutočnosti nepamätajú o nič viac než ja? Tipujem, že si iní ľudia niektoré veci pamätajú lepšie. Mne pripadá štruktorované vedecké poznanie dosť ľahké na zapamätenie; ale nespojitý chaos každodenného života sa mi veľmi rýchlo rozmazáva.

Viem, *prečo* sa niektoré veci v mojom živote stali – kauzálnu štruktúru si viem zapamätať. Ale niekedy je ťažké si spomenúť ešte aj *v akom poradí* sa mi isté veci stali, tobôž v ktorom roku.

Nie som si istý, či som čítal knihu *Teória pravdepodobnosti: Logika vedy* od E. T. Jaynesa predtým alebo potom, čo som si uvedomil rozsah svojej vlastnej hlúposti a uvedomil som si, že čelím problému pre dospelých.

Ale bola to *Teória pravdepodobnosti*, čo dosiahlo ten trik. Bola tu teória pravdepodobnosti, vysvetlená nie ako chytrý nástroj, ale ako *Pravidlá*, ktoré nemožno porušiť pod hrozbou paradoxu. Ak ste sa pokúsili tieto Pravidlá aproximovať, pretože boli príliš výpočtovo náročné na priame použitie, potom bez ohľadu na to, aký nevyhnutný bol tento kompromis, vo výsledku ste robili niečo menej než optimálne. Jaynes robil svoje výpočty rôznymi spôsobmi, aby ukázal, že keď používate legitímne metódy, vždy vyjdú rovnaké výsledky; a ukazoval rôzne odpovede, na ktoré prišli druhí, a vystopoval ten nelegitímny krok. Paradoxy nemohli pri jeho presnosti existovať. Nebola *nejaká* odpoveď, ale *odpoveď*.

A tak – keď som sa obzrel späť na svoje chyby, a všetky tie *nejaké odpovede*, ktoré ma doviedli k paradoxu a zdeseniu – uvedomil som si, že existuje úroveň vyššia než je tá moja.

Už som si viac nedokázal predstavovať, že skúsim zostrojiť UI na základe nejasných odpovedí – ako boli tie nejaké odpovede, s ktorými som prišiel predtým – a prežijem túto výzvu.

Pozrel som sa na rádoby expertov na VUI, s ktorými som skúšal debatovať o Priateľskej UI, a na rôzne sny o Priateľskosti, ktoré mali. (Často sformulované spontánne ako odpoveď na moju otázku!) Tak ako štatistické metódy frekventistov, žiadni dvaja medzi nimi sa nezhodli jeden s druhým. Keďže som túto tému naozaj študoval na plný úväzok niekoľko rokov, vedel som niečo o problémoch, na ktoré by ich nádejné plány narazili. A videl som, že keď poviete: „Nevidím, ako by toto mohlo zlyhať“, slovo „nevidím“ bolo iba odrazom vašej vlastnej nevedomosti. Videl som, že keby som si stanovil podobnú latku „vyzerá to ako dobrý nápad“, bol by som stratený. (Podobne ako frekventisti, keď vymýšľajú úžasné nové štatistické výpočty, ktoré vyzerajú ako dobrý nápad.)

Ale ak nemôžete urobiť to, čo vyzerá ako dobrý nápad – ak nemôžete urobiť to, čo si neviete predstaviť, že by zlyhalo – čo potom môžete robiť?

Pripadalo mi, že by to vyžadovalo niečo na úrovni, ako je Jaynes – nie: *tu je môj skvelý nápad*, ale skôr: *tu je jediný správny spôsob, ako sa to má robiť (a dôvod prečo)* – aby sa dalo zápasiť s problémom pre dospelých a prežiť. Keby som vo svojej oblasti dosiahol rovnaký stupeň majstrovstva ako Jaynes

dosiahol v teórii pravdepodobnosti, potom sa prinajmenšom dalo *predstaviť*, že by som mohol skúsiť zostaviť Priateľskú UI a prežiť to.

Mysľou pri prebehli slová:

> *Nerobte nič preto, lebo je spravodlivé, alebo chvályhodné, alebo vznešené robiť to tak; nerobte nič preto, lebo vyzerá dobré robiť to tak; robte iba to, čo musíte robiť, a čo nemôžete robiť žiadnym iným spôsobom.*[292]

Robiť to, čo vyzeralo dobré tak robiť, ma iba zviedlo z cesty.

Tak som vyhlásil úplné zastavenie.

A rozhodol som sa, že odteraz sa budem riadiť stratégiou, ktorá by ma bola zastavila, keby som sa ňou riadil pred rokmi: Nastaviť svojim dizajnou PUI vyššiu latku nerobenia toho, čo vyzerá ako dobrý nápad, ale iba toho, čomu rozumiem dostatočne hlboko na to, aby som videl, že sa to nedá urobiť nijako inak.

Všetky moje teórie, do ktorých som tak veľa investoval, nedosahovali túto latku; ani k nej neboli blízko; ani len neboli na ceste vedúcej k tejto latke; tak som ich vyhodil z okna.

Pustil som sa do štúdia teórie pravdepodobnosti a teórie rozhodovania, snažil som sa ich rozšíriť, aby zahrnuli také veci ako reflexivita a sebamodifikácia.

Ak si správne spomínam, už som v tomto bode začal vidieť poznávanie ako prejav bayesovskej štruktúry, čo je tiež veľká časť toho, čo označujem za svoje bayesovské osvietenie – ale o tom som už hovoril. A prišlo aj moje naturalistické prebudenie, o ktorom už som hovoril. A moje uvedomenie, že Tradičná Rozumnosť nie je dosť prísna, takže som sa vo veciach ľudskej rozumnosti začal inšpirovať teóriou pravdepodobnosti a kognitívnou psychológiou.

Ale keď dáte všetky tieto veci dokopy, potom je to viacmenej príbeh o mojom bayesovskom osvietení.

Život zvyčajne nemá pekné hranice. Príbeh pokračuje ďalej.

Počas štúdia Judeu Pearla som si napríklad uvedomil, že presnosť vám môže ušetriť čas. Predtým by som bol sám vemoval myšlienky nemonotónnej logike – vtedy, keď som ešte bol v režime „hľadanie pekných nástrojov a algoritmov“. Pri čítaní knihy *Pravdepodobnostné uvažovanie v inteligentných systémoch: Siete dôveryhodného odvodzovania*[293] som si vedel predstaviť, koľko času by som premárnil na ad-hoc systémy a špeciálne prípady, keby som nepoznal tento kľúč. „Robte iba to, čo musíte robiť, a čo by ste nemohli robiť žiadnym iným spôsobom,“ sa preložilo na časové úspory, nie na záchranu stratených mesiacov, ale na záchranu stratených kariér.

A tak som si uvedomil, že iba vdaka nastaveniu si tento vyššej latky presnosti som začal *naozaj* rozmýšľať o pomerne veľkom množstve dôležitých vecí. Povedať niečo presne je ťažké – nie je to vôbec to isté ako povedať niečo formálne, alebo vymyslieť si nejakú novú logiku, ktorú hodíme na tento problém. Mnohí sa vyhýbajú tomuto nepohodliu, pretože ľudia sú leniví, a tak povedia: „To sa nedá“ alebo „To by trvalo príliš dlho“, hoci sa nikdy neozaj nepokúšali ani päť minút. Ale ak si nenastavíte *nepohodlne* vysokú latku, potom si odpustíte hocičo. Ťažkým problémom je už len nájsť latku dosť vysokú na to, aby ste naozaj začali myslieť! Môže vám pripadať vyčerpávajúce držať sa latky matematického dôkazu, kde každý jeden krok musí byť správny a jediný zlý krok vás môže odviesť hocikam. Ale inak by ste nesledovali všetky tie tenké tóny nesúladu, ktoré nakoniec povedú k celkom novým stránkam, na ktoré ste predtým nikdy nepomysleli.

V týchto dňoch sa teda už tak veľmi nesťažujem na hrdinské bremeno nepohodlnosti, ktorú treba, aby ste si postavili vysokú latku. Môže to aj šetriť čas; a vlastne je to viacmenej predpoklad, aby ste sa donútili o danom probléme vôbec rozmýšľať.

A aj toto treba považovať za časť môjho „bayesovského osvietenia“ - uvedomenie si, že to má svoje výhody, nielen tresty.

Ale samozrejme príbeh pokračuje. Život je taký, aspoň tie časti, ktoré si pamätám.

---

292 Le Guin, *The Farthest Shore*.

293 Pearl, *Probabilistic Reasoning in Intelligent Systems*.

Ak som sa z tejto histórie niečo naučil, je to že povedať „Joj“ je niečo, na čo sa treba tešiť. Iste, predstava, že poviete „Joj“ v budúcnosti znamená, že vaše *terajšie ja* je slintajúci imbecil, ktorého slová vašej budúce ja nebude schopné prečítať bez toho, že by sa pritom skrúcalo. Lenže povedať „Joj“ v budúcnosti znamená aj to, že v tej budúcnosti nadobudneté nové Jediovské schopnosti, o ktorých existencii vaše terajšie ja ani nesníva. Budete sa cítiť zahanbene, ale zároveň *živo*. Uvedomíť si, že vaše mladšie ja bolo úplný blbec, znamená, že aj keď už máte po dvadsiatke, ešte stále nie ste za svojím vrcholom. To je teda dôvod dúfať, že si moje budúce ja uvedomí, že som slintajúci imbecil: Mám síce *v pláne* vyriešiť svoje problémy svojimi súčasnými schopnosťami, ale nejaké Jediovské schopnosti by mi iste prišli vhod.

Ten krik hrôzy a hanby, to je zvuk, ktorý racionalisti vydávajú, keď postupujú na vyššiu úroveň. Niekedy sa bojím, že nepostupujem tak rýchlo ako kedysi, a neviem, či je to preto, že už konečne dochádzam veciam na koreň, alebo preto, lebo neuróny v mojom mozgu pomaly umierajú.

S pozdravom, Eliezer$_{2008}$.

* →

→ http://lesswrong.com/lw/ul/my_bayesian_enlightenment/

Y: Pustiť sa do ťažkých vecí

304. Tsuyoku naritai! (Chcem sa stať silnejším)

Ortodoxný judaizmus má príslovie: „Predchádzajúca generácia je voči nasledujúcej, ako anjeli voči ľuďom; nasledujúca generácia je voči predchádzajúcej, ako somáre voči ľuďom.“ Vychádza to z ortodoxnej židovskej viery, že na hore Sinaj dal Boh Mojžišovi všetky židovské zákony. Nemôžete predsa urobiť žiaden experiment, ktorým by ste získali nové halachické poznanie; jediný spôsob, ako ho môžete mať, je keď vám to niekto povie (kto to počul od niekoho iného, kto to počul od Boha). Keďže neexistuje žiaden nový zdroj informácií, informácie sa môžu pri prenose z generácie na generáciu iba zhoršovať.

Moderní rabíni preto nemajú dovolené odporovať dávnym rabínom. Plazivé veci zvyčajne nie sú košer, ale je prípustné zjesť červíka nájdeného v jablku – dávni rabíni verili, že červík vznikol spontánne vnútri jablka, a preto je súčasťou jablka. Moderný rabín nemôže povedať: „No viete, dávni rabíni vedeli figu o biológii. Neplatí!“ Moderný rabín nemôže žiadnym spôsobom spoznať halachický princíp, ktorí dávni rabíni nepoznali, pretože ako by mu dávni rabíni mohli odovzdať túto informáciu z hory Sinaj? Poznanie sa odvodzuje z autority, a preto sa postupom času môže iba strácať, nie získavať.

Keď som sa prvýkrát stretol s týmto príslovím o anjeloch a somároch na (náboženskej) základnej škole, nebol som dosť starý na to, aby som bol zrelým ateistom, ale aj tak som si pomyslel: „Tóra každou generáciou stráca poznanie. Veda každou generáciou získava poznanie. Bez ohľadu na to, kde začali, skôr či neskôr musí veda predbehnúť Tóru.“

Najdôležitejšia vec je, aby existoval pokrok. Dokiaľ sa hýbete vpred, dosiahnete svoj cieľ; keď sa však prestanete hýbať, nikdy ho nedosiahnete.

Tsuyoku naritai [„cüjokü naritaj“] je po japonsky. *Tsuyoku* znamená „silný“, *naru* znamená „stávanie sa“ a tvar *naritai* znamená „chcieť sa stať“. Dohromady to znamená: „Chcem sa stať silnejším“ a vyjadruje to cítenie stvárnené v japonských dielach omnoho intenzívnejšie než v akejkoľvek západnej literatúre, ktorú som čítal. Môžete to povedať, aby ste vyjadrili svoje odhodlanie stať sa profesionálnym hráčom Go... alebo potom, čo prehráte dôležitú partiu, ale nevzdávate sa... alebo potom, čo vyhráte dôležitú partiu, ale ešte nie ste hráč s deviatym danom... alebo potom, čo sa stanete najlepším hráčom Go všetkých čias, no stále si myslíte, že sa ešte môžete zlepšiť. To je *tsuyoku naritai*, vôla prekonávať.

Tsuyoku naritai je hlavnou silou za mojím článkom Správne použitie pokory, v ktorom dávam študenta, ktorý si pokorne urobí skúšku správnosti na písomke z matematiky, do protikladu k študentovi, ktorý pokorne povie: „Ale ako môžeme niečo naozaj vedieť? Bez ohľadu na to, koľko skúšok správnosti urobím, nikdy si nemôžem byť absolútne istý.“ Ten študent, ktorý robí skúšku správnosti, *chce byť silnejším*; na svoju možnú chybu reaguje robením, čo môže, aby chybu napravil, nie rezignáciou.

Každoročne na Yom Kippur ortodoxný žid cituje litániu, ktorá začína: *Ashamnu, bagadnu, gazalnu, dibarnu dofi...* a pokračuje cez celú hebrejskú abecedu: *Konali sme hanebne, zradili sme, kradli sme, ohovárali sme...*

Pri vyslovení každého slova sa udriete na srdce na znak ľútosti. Neexistuje žiadna výnimka, že ak sa vám podarí nekradnúť po celý rok, môžete vynechať slovo *gazalnu* a udrieť sa o jeden raz menej. To by narušilo spoločnú atmosféru Yom Kippuru, ktorá je o *spovedaní sa* z hriechov – nie o *vyhýbaní sa* hriechom, aby ste sa mohli menej spovedať.

Rovnako, *Ashamnu* nekončí: „Ale to bolo tento rok, a ďalší rok budem lepší.“

Ashamnu sa pozoruhodne podobá na predstavu, že cesta k rozumnosti je udierať päsťou na srdce a hovoriť: „Všetci sme ovplyvňovaní sklonmi, všetci sme iracionálni, všetci sme nedostatočne informovaní, všetci sme príliš sebavedomí, všetci sme zle kalibrovaní...“

Dobre. Teraz mi ale povedzte, ako sa plánujete stať *menej* ovplyvňovaní sklonmi, *menej* iracionálni, *viac* informovaní, *menej* prehnane sebavedomí, *lepšie* kalibrovaní.

Je taký starý židovský vtip: Počas Yom Kippuru zachváti rabína náhla vlna pocitu viny, ľahne si na zem a vykríkne: „Bože, pred tebou nie som nič!“ Kantora podobne zachváti pocit viny a vykríkne: „Bože,

pred tebou nie som nič!“ Keď to vidí vrátnik na konci synagógy, ľahne si na zem a vykríkne: „Bože, pred tebou nie som nič!“ A rabín drgne do kantora a zašepká: „Aha, kto sa tu považuje za nič.“

Nebuďte hrdí na svoju spoveď, že aj vy sa mýlite; nevystatujte sa uvedomovaním si vlastných omylov. Podobne nebuďte hrdí na spovedanie sa z nevedomosti; lebo ak sa vám vaša nevedomosť stane zdrojom hrdosti, možno nebudete ochotní sa tejto nevedomosti vzdať, keď vám indície zaklopú na dvere. Podobne je to s našimi chybami – nemali by sme sa tešiť z toho, akí uvedomelí sme, že sa z nich spovedáme; dôvodom na radosť je, keď toho máme na spovedanie o čosi menej.

V opačnom prípade, keď jeden z nás príde a predloží plán, ako daný omyl *opraviť*, budeme naňho vrčať: „Vari si myslíš, že si lepší než my?“ Smutne pokrútime hlavami a povieme: „Zrejme toho o sebe málo vieš.“

Nespovedajte sa mi, že máte tie isté chyby ako ja, pokiaľ mi zároveň nepoviete, čo s tým plánujete urobiť. Aj potom vám ešte zostane kopec chýb, ale o to nejde: dôležité je *zlepšovať sa*, ísť stále vpred, urobiť ďalší krok. *Tsuyoku naritai!*

*→

305. Tsuyoku verzus rovnostársky inštinkt

Tlupy lovcov a zberačov sú zvyčajne veľmi rovnostárske (prinajmenšom ak ste muž) – všemocní kmeňoví náčelníci sa zvyčajne vyskytujú v poľnohospodárskych spoločnostiach, zriedkavo v pravekých. U väčšiny tlúp lovcov a zberačov si lovec, ktorý prinesie mimoriadnu korisť, dá veľmi záležať na tom, aby svoj úspech bagatelizoval, aby sa vyhol závisti.

Možno, ak začínate ako podpriemerný, sa môžete zdokonaľovať aj bez odvahy predbehnúť dav. Ale skôr či neskôr, ak sa snažíte o to najlepšie, čo je vo vašich silách, sa začnete pokúšať o niečo nadpriemerné.

Ak si nedokážete priznať, že ste na tom v niečom lepšie než ostatní – alebo ak sa hanbíte za to, že na tom chcete byť lepšie než ostatní – potom bude medián navždy vašou betónovou stenou, miestom, kde prestanete ísť vpred. A čo s ľuďmi, ktorí sú podpriemerní? Odvážite sa povedať, že chcete na tom byť lepšie než oni? Aká pýcha!

Možno nie je zdravé cítiť sa hrdý na to, že ste na tom lepšie než niekto iný. Pre mňa osobne je to užitočná motivácia, napriek mojim zásadám, a každá užitočná motivácia, ktorú viem zohnať, sa mi hodí. Možno je tento druh súťaživosti hrou s nulovým súčtom, ale takisto je aj Go; neznamená to, že by sme mali zrušiť túto ľudskú činnosť, keď ľuďom pripadá zábavná a vedie k niečomu zaujímavému.

Každopádne, určite nie je zdravé *hanbiť sa* za to, že ste na tom lepšie.

A okrem toho, život sa neznámkuje podľa stupnice. Vôľa prekonávať nemá žiaden bod, za ktorým skončí a premení sa na vôľu robiť veci horšie; preteky, ktoré nemajú cieľovú pásku, nemajú ani zlaté a strieborné medaile. Jednoducho bežte tak rýchlo, ako môžete, a netrápte sa tým, že možno predbehnete iných bežcov. (Ale varujem vás: Ak sa odmietnete trápiť touto možnosťou, jedného dňa ich možno naozaj predbehnete. Ak ignorujete tieto dôsledky, môžu sa vám prihodiť.)

Skôr či neskôr, ak vaša cesta vedie správne, sa rozhodnete zmenšiť chybu, ktorú väčšina ľudí nezmenšila. Skôr či neskôr, ak vaše úsilie prináša nejaké ovocie, zistíte, že máte na spovedanie menej hriechov.

Možno zistíte, že je múdrou reakciou bagatelizovať svoje úspechy, aj keď sa vám darí. Ľudia vám možno odpustia gól, ale nie tanec pred bránkou. Iste zistíte, že je rýchlejšie, ľahšie a pohodlnejšie verejne popierať svoju hodnotu, predstierať, že ste len rovnaký hriešnik ako každý iný. Samozrejme pod podmienkou, že každý vie, že to nie je pravda. Môže byť zábava verejne ukazovať svoju skromnosť, pokiaľ každý vie, že je toho veľmi veľa, na čo ste skromný.

Ale nedovoľte, aby sa toto stalo cieľom vášho putovania. Aj keď si to budete iba šepkať, aj tak si šepkajte: *Tsuyoku, tsuyoku!* Silnejší, silnejší!

→ http://lesswrong.com/lw/h8/tsuyoku_naritai_i_want_to_become_stronger/

A potom si vytýčte vyšší cieľ. To je skutočný význam uvedomovania si, že ešte stále máte chyby (aj keď o čosi menej). Znamená to vždy siahať vyššie, bez hanby.

Tsuyoku naritai! Vždy pobežím tak rýchlo, ako môžem, a aj keď sa dostanem dopredu, stále budem bežať; a jedného dňa ma niekto predbehne, ale aj keď zaostanem, stále budem bežať tak rýchlo, ako môžem.

*→

306. Pokúsiť sa pokúsiť

Nie! Nepokúšaj sa! Urob, alebo neurob. Neexistuje žiadne pokúsim sa.

--Yoda

Pred rokmi som si myslel, že toto je iba ďalší príklad Hlbokej Múdrosti, ktorý je v skutočnosti dosť hlúpy. ÚSPECH nie je primitívna akcia. Nemôžete sa skrátka *rozhodnúť*, že vyhráte tým, že sa dostatočne silno rozhodnete. Neexistuje žiaden plán, ktorý funguje s pravdepodobnosťou 1.

Lenže Yoda bol múdrejší, než som si prvýkrát uvedomil.

Prvá základná technika epistemológie – nie je hlboká, ale je lacná – je rozlišovať citáciu a referenta. Hovoriť o snehu nie je to isté ako hovoriť o „snehu“. Keď použijem slovo „sneh“ bez úvodzoviek, chcem hovoriť o snehu; a keď použijem slovo „sneh“ s úvodzovkami, chcem hovoriť o slove „sneh“. Musíte vstúpiť do osobitného režimu, režimu citácií, aby ste hovorili o svojich názoroch. Štandardne totiž hovoríme o skutočnosti.

Ak niekto povie: „idem prepnúť ten prepínač,“ potom tým štandardne myslí, že ide prepnúť ten prepínač. Ide zostaviť plán, ktorý sľubuje, že v dôsledku svojich činov vedie k cieľovému stavu prepnutého prepínača; a potom ide vykonať tento plán.

Žiaden plán neuspeje s nekonečnou istotou. Takže štandardne, keď hovoríte o rozhodnutí dosiahnuť nejaký cieľ, nenaznačujete, že váš plán presne a dokonale vedie *iba* k tejto možnosti. Ale keď poviete: „Idem prepnúť ten prepínač“, *snažíte* sa iba prepnúť ten prepínač – *nesnažíte* sa dosiahnuť pravdepodobnosť 97,2 %, že prepnete prepínač.

Čo teda znamená, keď niekto povie: „Idem sa *pokúsiť* prepnúť ten prepínač?“

Nuž, *v bežnej reči* „Idem prepnúť ten prepínač“ a „Idem sa pokúsiť prepnúť ten prepínač“ znamená viacmenej to isté, akurát že to druhé vyjadruje možnosť zlyhania. To je dôvod, prečo ma pôvodne pohoršilo, že Yoda vyzerá, akoby popieral túto možnosť. Ale vydržte ešte chvíľu.

Mnoho zo životných výziev spočíva v tom, že sa pridŕžame dostatočne vysokého štandardu. O tomto princípe možno neskôr poviem viac, pretože je to objektív, ktorým môžete vidieť mnohé, aj keď nie všetky, osobné dilemy - „Aký štandard od seba očakávam? Je dosť vysoký?“

Ak teda mnohé životné zlyhania spočívajú v tom, že ste si určili príliš nízky štandard, mali by ste sa vyhýbať očakávaniu od seba príliš málo – stanovovaniu si cieľov, ktoré sa príliš ľahko plnia.

Často tam, kde *uspieť* v niečom je veľmi náročné, *pokúsiť sa* to robiť je omnoho ľahšie.

Čo je ľahšie – vybudovať úspešnú firmu, alebo pokúsiť sa vybudovať úspešnú firmu? Zarobiť milión dolárov, alebo pokúsiť sa zarobiť milión dolárov?

Ak teda: „Idem prepnúť ten prepínač“ štandardne znamená, že sa idete pokúsiť prepnúť ten prepínač – čiže, že idete zostaviť plán, ktorý sľubuje, že vedie k stavu prepnutého prepínača, možno nie s pravdepodobnosťou 1, ale s najvyššou pravdepodobnosťou, akú dokážete...

...potom: „Idem sa ‚pokúsiť prepnúť‘ ten prepínač“ znamená, že sa idete pokúsiť „pokúsiť sa prepnúť ten prepínač“, čiže, že sa idete pokúsiť dosiahnuť cieľový stav „mám plán, ktorý by mohol prepnúť ten prepínač“.

Takže, keby sme teraz hovorili o sebamodifikujúcej sa UI, transformácia, ktorú sme práve urobili, by mala skončiť v reflexívnej rovnováhe – UI plánuje svoje plánovacie operácie.

→ http://lesswrong.com/lw/h9/tsuyoku_vs_the_egalitarian_instinct/

Ale keď máme do činenia s ľuďmi, *byť spokojný s tým, že máme plán*, nie je vôbec také ako *byť spokojný s úspechom*. Tá časť, kde má plán maximalizovať vašu pravdepodobnosť úspechu, sa cestou stratí. Je omnoho jednoduchšie presvedčiť sami seba, že „maximalizujeme našu pravdepodobnosť úspechu“, než presvedčiť sami seba, že sme uspeli.

Takmer hocijaké úsilie nám poslúži, aby sme sa presvedčili, že sme sa „snažili, ako sa len dalo“, ak snažiť sa, ako sa len dá, je všetko, o čo sa pokúšame.

> Pýtal si sa, čo by si mohol robiť v tých veľkých udalostiach, ktoré dnes prebiehajú, a zistil si, že by si nemohol urobiť nič. Ale to preto, lebo tvoje trápenie spôsobilo, že si si otázku postavil nesprávnym spôsobom... Namiesto pýtania sa, čo by si mohol urobiť, si sa mal pýtať, čo treba urobiť.
>
> --Steven Brust, Cesty mŕtvych[294]

Keď sa opýtate: „Čo môžem urobiť?“ snažíte urobiť, čo sa dá? Čo je to, čo sa dá? Je to to, čo môžete urobiť bez najmenšieho nepohodlia. Je to to, čo môžete urobiť s peniazmi vo vašom vrecku, mínus koľko potrebujete na svoj zvyčajný obed. Čo môžete urobiť s týmito prostriedkami, vám možno nedáva veľmi dobré šance na víťazstvo. Ale je to to, „čo sa dalo urobiť“, takže sa vaše konanie dá obhájiť,však?

Ale čo *treba* urobiť? Možno to, čo *treba* urobiť, vyžaduje trojnásobok vašich životných úspor, a musíte to urobiť alebo zlyhať.

Takže snažiť sa mať „maximalizovanú svoju pravdepodobnosť úspechu“ - na rozdiel od snahy uspieť – je omnoho nižšia latka. Môžete „maximalizovať svoju pravdepodobnosť úspechu“ s použitím iba peňazí vo vašom vrecku, dokiaľ nežiadate, aby ste naozaj *vyhrali*.

Chcete sa pokúsiť vyhrať milión dolárov? Kúpte si žreb z lotérie. Vaše šance vyhrať asi nebudú veľmi dobré, ale pokúsili ste sa, a pokúsiť sa bolo to, čo ste chceli. V skutočnosti ste urobili to *najlepšie*, čo sa dalo, pretože vám po kúpení obeda zostal vo vrecku iba jeden dolár. Maximalizovať šance dosiahnutia cieľa pri použití dostupných zdrojov: nie je toto inteligencia?

Jedine keď chcete, nado všetko ostatné, *naozaj* prepnúť ten prepínač – bez citácie a bez ceny útechy za púhy pokus – vtedy *naozaj* vynaložíte všetko úsilie na to, aby ste *naozaj* maximalizovali pravdepodobnosť.

Ale ak všetko, čo chcete, je „maximalizovať pravdepodobnosť úspechu pri vynaložení dostupných zdrojov“, potom je tá najjednoduchšia vec na svete presvedčiť sa, že ste hotoví. Celkom prvý plán, na ktorý natrafíte, vám poslúži celkom dobre ako „maximalizácia“ - ak treba, dokážete vytvoriť horšiu alternatívu, aby ste dokázali jeho optimálnosť. A ľubovoľné malé zdroje, ktoré sa vám bude chcieť vynaložiť, budú to, čo je „dostupné“. Nezabudnite si sami zablahoželať, že ste do toho vložili 100 %!

Nepokúšajte sa urobiť to, čo sa dá. Vyhrajte, alebo zlyhajte. Neexistuje žiadne čo sa dá.

*→

307. Použi námahu, Luke

> Kde je vôľa zlyhať, prekážky sa nájdu.
>
> --John McCarthy

Prvýkrát som pozeral *Hviezdne vojny* IV-VI, keď som bol veľmi malý. Možno sedem rokov, alebo deväť? Moja pamäť bola teda rozmazaná, ale spomínal som si, že Luke Skywalker bol, veď viete, ten cool Jedi chlapík.

Predstavte si môju hrôzu a sklamanie, keď som tú ságu sledoval znovu, o niekoľko rokov neskôr, a zistil som, že Luke je umrnčaný puberťák.

---

294 Steven Brust, *The Paths of the Dead*, Vol. 1 of The Viscount of Adrilankha (Tor Books, 2002).

→ http://lesswrong.com/lw/uh/trying_to_try/

Spomínam to preto, lebo som si včera vyhľadal na YouTube zdroj Yodovho citátu: „Rob, alebo nerob. Neexistuje žiadne pokúsim sa.“

Ach. Môj. Cthulhu.

Popri spomínanom klipe z YouTube vám predstavím málo známu scénu z natáčania, v ktorej sa režisér a autor George Lucas háda s Markom Hamillom, ktorý hral Luka Skywalkera:

Luke: Dobre, pokúsim sa.

Yoda: Nie! Nepokúšaj sa. Urob. Alebo neurob. Neexistuje žiadne pokúsim sa.

Luke zdvihne ruku a X-wing sa pomaly začne dvíhať z vody – Yodovi sa rozšíria oči – ale potom sa loď opäť potopí.

Mark Hamill: „Hej, George...“

George Lucas: „Čo zase?“

Mark: „Takže... podľa scenára teraz poviem: ‚Nemôžem. Je veľmi veľká.‘“

George: „Presne tak.“

Mark: „Nemal by to Luke trebárs skúsiť po druhýkrát?“

George: „Nie. Luke sa vzdá a sadne si vedľa Yodu...“

Mark: „Toto je ten hrdina, ktorý má poraziť Impérium? Pozri, jedna vec bola, keď bol na začiatku umrnčaný puberťák, ale teraz je v tréningu u Jediov. V predchádzajúcom filme vyhodil do vzduchu Hviezdu Smrti. Luke by už mal mať trochu chrbtovej kosti.“

George: „Nie. Vzdáš sa. A potom ťa Yoda bude chvíľu poúčať a potom povieš: ‚Chceš nemožné.‘ Dokážeš si to zapamätať?“

Mark: „*Nemožné?* To akože urobil formálny výpočet a došiel k matematickému dôkazu? Ten X-wing sa predsa už začal dvíhať z močiara! To je predvedenie uskutočniteľnosti priamo na mieste! Luke to na sekundu nezvládne a loď sa opäť potopí – a teraz hovorí, že je to *nemožné*? Nehovoriac o tom, že Yoda, ktorý je doslova osemsto rokov odborníkom na túto oblasť, mu práve povedal, že by sa to malo dať...“

George: „A potom odídeš preč.“

Mark: „Veď je to sakra jeho *vesmírna loď!* Keď ju nechá v močiari, zakysne na Dagobahu na celý zvyšok svojho biedneho života! Nemôže len tak odísť preč! Pozri, čo keby sme dali prestrih na ďalšiu scénu so slovami ‚o mesiac neskôr‘, a Luke bude stále vyčerpane stáť pred močiarom a pokúšať sa po tisícikrát zdvihnúť svoju loď...“

George: „Nie.“

Mark: „Dobre! Tak ukážeme západ a východ slnka, ako tam stojí s vystretou rukou, namáha sa, a *potom* povie: ‚Je to nemožné.‘ Hoci v skutočnosti by sa o to mal pokúsiť znova, keď si poriadne odpočinie...“

George: „Nie.“

Mark: „*Päť poondených minút!* Päť poondených minút predtým, než sa vzdá!“

George: „Nebudem zdržiavať dej na päť minút, kým sa bude X-wing hojdať v močiari ako hračka vo vani.“

Mark: „V mene sladkých kandizovaných zemiačikov! Keby takýto patetický úbožiak dokázal ovládnuť Silu, potom by ju používal každý v celej galaxii! Ľudia by sa stali Jediami, pretože by to bolo ľahšie než chodiť na strednú školu.“

George: „Pozri, ty si herec. Nechaj mňa rozprávať príbeh. Skrátka prednes svoje repliky, a nech to znie, že to tak myslíš.“

Mark: „Toto diváci nezožerú.“

George: „Dôveruj mi, zožerú.“

Mark: „Zdvihnú sa a odídu z kina.“

George: „Budú tam sedieť a prikyvovať a nevšimnú si nič nezvyčajné. Pozri, ty nechápeš ľudskú povahu. Ľudia by sa nesnažili ani päť minút, než by sa vzdali, keby bol v stávke osud celého ľudstva.“

*→

308. O robení nemožného

„Buď vytrvalý.“ Je to jedna rada, ktorú dostanete od mnohých veľmi úspešných ľudí v mnohých oblastiach. Najprv som tomu vôbec nerozumel.

Najprv som si myslel, že „vytrvalosť“ znamená pracovať 14 hodín denne. Vyzerá to, že existujú ľudia, ktorí dokážu 10 hodín pracovať v nejakej technickej práci, a potom, vo chvíľkach medzi jedením a spánkom a chodením na záchod využívajú nezaplnený voľný čas na písanie knihy. Ja nie som jeden z nich – aj teraz zraňuje moju hrdosť, keď to priznávam. Pracujem na niečom dôležitom; nemal by môj mozog byť ochotný makať 14 hodín denne? No nie je. Keď začne byť príliš ťažké pokračovať v práci, prestanem, a idem si niečo prečítať alebo pozrieť. Preto som si celé roky myslel, že mi celkom chýba cnosť „vytrvalosť“.

V súlade s ľudskom povahou by si Eliezer$_{1998}$ bol myslel veci ako: „Dôležitý je výstup, nie vstup.“ Alebo: „Lenivosť je tiež cnosť – vedie nás preč od nefungujúcich metód, aby sme mysleli na lepšie spôsoby.“ Alebo: „Darí sa mi lepšie než iným ľuďom, ktorí pracujú dlhšie. Možno je pri tvorivej práci váš momentálny *vrcholný* výkon dôležitejší než pracovať 16 hodín denne.“ Možno známych vedcov zlákala Hlboká Múdrosť hovorenia, že „tvrdá práca je cnosť“, pretože by bolo príliš hrozné, keby znamenala menej než inteligencia?

Nerozumel som cnosti vytrvalosti, dokiaľ som sa neobzrel späť na svoju púť po UI a neuvedomil som si, že som preceňoval náročnosť takmer každého jedného dôležitého problému.

Znie to šialene, však? Ale vydržte so mnou.

Keď som sa prvýkrát pokúšal o výzvu UI, uvažoval som o 40-ročnej časovej škále, o projektoch Manhattan, o celoplanetárnych výpočtových sieťach, o miliónoch programátorov, a možno aj o vylepšených ľuďoch.

Toto je bežný režim zlyhania vo futurizme UI, o ktorom možno napíšem neskôr; spočíva v skoku z „neviem, ako toto vyriešiť“ na „predstavujem si, že na to použijeme niečo naozaj veľké“. Niečo dosť veľké na to, že keď si to predstavíte, vaša predstavivosť vytvorí pocit dosť silného dojmu, ktorý je porovnateľný s problémom. (V skupine UI je momentálne človek, ktorý hovorí, že UI bude stáť biliardu dolárov – že nemôžeme mať UI *bez* minutia biliardy dolárov, ale *mohli* by sme mať UI hocikedy, keby sme na ňu minuli biliardu dolárov.) To vám potom dovoľuje si predstavovať, že viete, ako vyriešiť UI, a nemusíte sa snažiť splniť tú očividne nemožnú podmienku, že musíte *rozumieť, čo je to inteligencia*.

Na začiatku som teda urobil rovnakú chybu: Nerozumel som, čo je to inteligencia, tak som si predstavil, že na ten problém použijem projekt Manhattan.

Ale keď som si spočítal, že na planéte zomrie 55 miliónov ľudí ročne, alebo 150 000 denne, neotočil som sa nezačal utekať preč od veľkého strašného problému ako vyplašený zajac. Namiesto toho som začal skúšať zistiť, aký druh projektu UI by nás tam mohol dostať najrýchlejšie. Keby som dokázal urobiť, aby explózia inteligencie nastala o jednu hodinu skôr, bola by to rozumná návratnosť investovania celej kariéry pred explóziou. (V tom bode som ešte neuvažoval nad existenciálnym rizikom alebo Priateľskou UI.)

Takže som nezačal utekať preč od veľkého strašného problému ako vyplašený zajac, ale zostal som, aby som videl, či je niečo, čo by som mohol urobiť.

Zaujímavý historický fakt: V roku 1998 som napísal dlhé pojednanie navrhujúce, ako by sa mala vytvoriť sebazdokonaľujúca alebo „zárodočná“ UI (mal som česť vytvoriť tento pojem). Brian Atkins, ktorý sa neskôr stal zakladajúcim sponzorom Výskumného ústavu strojovej inteligencie, práve predal Hypermart firme Go2Net. Brian mi napísal e-mail, v ktorom sa opýtal, či tento projekt UI, ktorý opisujem, je niečo, čo by tím rozumnej veľkosti mohol naozaj ísť *urobiť*. „Nie,“ povedal som, „to by

→ http://lesswrong.com/lw/ui/use_the_try_harder_luke/

vyžadovalo projekt Manhattan a tridsať rokov," takže sme chvíľu uvažovali, že namiesto toho urobíme novú dotcomovú firmu, aby sme zabezpečili financie na *skutočnú* prácu na UI...

O rok ale o dva neskôr, keď som počul o tejto novej veci „open source", mi pripadalo, že existuje nejaká predbežná vývojárska práca – nový počítačový jazyk a podobne – ktorú by mohla urobiť aj malá organizácia; a takto začal MIRI.

Táto stratégia bola, samozrejme, celá zle.

No napriek tomu som prešiel z: „Neexistuje nič, čo by som s tým mohol teraz urobiť" na: „Hm... možno existuje postupná cesta cez vývoj open source, keby tie počiatočné verzie boli užitočné pre dostatok ľudí."

Toto bolo vtedy na začiatku času, takže nehovorím, že niečo z toho je *dobrá myšlienka*. A z pohľadu toho, čo som si myslel, že sa pokúšam urobiť, rok tvorivého myslenia skrátil domnelú cestu: Problém vyzeral *o trochu menej nemožne*, než keď som sa k nemu priblížil po prvýkrát.

Zaujímavejší vzor je môj vstup do Priateľskej UI. Pôvodne som na niečím ako Priateľská UI ani nerozmýšľal – pretože bolo *očividne nemožné a zbytočné* klamať superinteligencii o tom, čo je správna vec.

Historicky som teda prešiel od *úplného ignorovania problému, ktorý bol „nemožný"* k *pusteniu sa do problému, ktorý bol iba extrémne ťažký*.

Prirodzene to zvýšilo moje celkové vyťaženie.

Podobná vec pri pokuse pochopiť inteligenciu na presnej úrovni. Pôvodne by som odpísal tento problém ako *nemožný*, čím by som ho odstránil zo svojho bremena. (Táto logika pri spätnom pohľade vyzerá dosť pomätene – Prírode nezáleží na tom, čo dokážete urobiť, keď píše požiadavky na váš projekt – ale dodnes vidím, ako sa o to ľudia v UI stále pokúšajú.) Stanoviť si latku presnosti znamenalo venovať tomu viac práce než som si pôvodne myslel, že treba. Ale zároveň to znamenalo riešiť problém, ktorý by som ešte nedávno zavrhol ako *celkom nemožný*.

Hoci *jednotlivé* problémy v UI začali postupne vyzerať menej hrozivo, celková hora, na ktorú treba vyliezť, sa zvyšovala – ako nám bežná múdrosť hovorí, že sa deje – keď sa problémy presúvajú zo zoznamu „nemožné" a umiestňujú do zoznamu „treba urobiť".

Začal som chápať, čo sa deje – čo „Buď vytrvalý!" naozaj znamená – v bode, kde som si všimol, ako ostatní ľudia z UI robia to isté: hovoria „Nemožné!" na problémy, ktoré vyzerajú vysoko riešiteľné – pomerne priamočiare, v porovnaní s inými. Ale boli to veci, ktoré *by* vyzerali omnoho zastrašujúcejšie v bode, keď som sa prvýkrát pustil do tohto problému.

A uvedomil som si, že slovo „nemožné" má dva významy:

1) Matematický dôkaz nemožnosti, závisiaci na zadaných axiómach;

2) „Nevidím žiaden spôsob, ako na to."

Netreba dodávať, že vždy, keď som použil slovo „nemožné", bolo to toho druhého typu.

Vždy, keď nerozumiete nejakej oblasti, mnohé problémy v tej oblasti vám budú pripadať nemožné, pretože keď sa opýtate svojho mozgu, aký je postup riešenia, vráti null. Lenže existujú iba tajomné otázky, nikdy nie tajomné odpovede. Ak strávite rok alebo dva prácou v danej oblasti, potom, *ak* ste sa nezasekli v niektorej zo slepých uličiek, a *ak* máte prirodzenú úroveň schopnosti potrebnú na to, aby ste urobili pokrok, budete tomu rozumieť lepšie. *Zdanlivá* ťažkosť problémov môže výrazne klesnúť. Nebude to také strašné, ako to bolo pre vaše začínajúce ja.

A toto je mimoriadne pravdepodobné pri ***mätúcich*** *problémoch, ktoré vyzerajú* ***najhrozivejšie***.

Odkedy máme nejakú predstavu o procesoch, pomocou ktorých hviezda horí, vieme, že nie je ľahké vybudovať hviezdu od nuly. Pretože rozumieme ozubeným kolesám, vieme dokázať, že žiadna zbierka ozubených kolies riadiacich sa zákonmi fyziky nemôže tvoriť stroj na večný pohyb. Toto nie sú dobré problémy na trénovanie si robenia nemožného.

Keď ste v nejakej oblasti *zmätení*, problémy v nej vám budú *pripadať* veľmi zastrašujúce a tajomné, a otázka pre váš mozog bude vracať celkový počet nula riešení. Ale neviete, koľko práce zostane potom, keď sa zmätok vyjasní. Rozpustiť ten zmätok môže byť samozrejme samo osebe veľmi ťažký problém. Ale slovo „nemožné" by sa v tomto spojení nemalo používať. Zmätok existuje na mape, nie v území.

Ak teda strávite niekoľko rokov prácou na nemožnom probléme, a dokážete sa vyhnúť slepým uličkám, alebo sa z nich vyškriabať von, a vaša prirodzená schopnosť je dosť vysoká na to, aby ste robili pokrok, potom, hľa, o pár rokov to už vôbec nemusí vyzerať také *nemožné*.

Ale ak niečo vyzerá nemožné, ani sa nepokúsite.

To je začarovaný kruh.

Keby som nebol v dostatočne posadnutom stave mysle, že pre mňa „štyridsať rokov a projekt Manhattan“ znamenalo akurát to, že by sme mali začať skôr, nebol by som sa pokúsil. Nebol by som sa prilepil na tento problém. A nedostal by som šancu stať sa menej zastrašeným.

Za bežných okolností nie som fanúšikom teórie, že protikladné skreslenia sa navzájom vykrátia, ale niekedy sa to šťastne podarí. Keby som *na začiatku* videl celú tú horu – keby som si na začiatku uvedomil, že problémom nie je zostaviť zárodok schopný sebazdokonaľovania, ale vytvoriť *dokázateľne správnu Priateľskú UI* – potom by som asi vybuchol a zhorel.

Ešte aj tak, časť pochopenia tých nadpriemerných vedcov, ktorí tvoria väčšinu výskumníkov VUI je uvedomiť si, že nie sú takí *posadnutí*, aby sa pustili do takmer nemožného problému, aj keď im to bude trvať 40 rokov. Väčšinou sú tam preto, lebo našli ten správny Kľúč k UI, ktorý im umožní vyriešiť tento problém *bez* takej ohromnej ťažkosti, za nejakých päť rokov.

Richard Hamming zvykol klásť svojim kolegom vedcom dve otázky: „Čo sú tie najdôležitejšie problémy v tvojej oblasti?“ a „Prečo na nich nerobíš?“

Často tie dôležité problémy vyzerajú Veľké, Hrozné, a Zastrašujúce. Nesľubujú vám 10 publikácií za rok. *Nesľubujú* vám vôbec žiaden pokrok. Možno nedostanete žiadnu odmenu ani keď na nich budete pracovať rok, alebo päť rokov, alebo desať rokov.

A nie je nezvyčajné, že tie najdôležitejšie problémy vo vašej oblasti sú nemožné. Práve preto nevidíte viac filozofov pracovať na redukcionistických dekompozíciách vedomia.

Pokúšať sa urobiť nemožné jednoznačne nie je pre každého. Výnimočný talent je iba podmienkou, aby ste si sadli ku stolu. Hracie žetóny sú roky vášho života. Ak vám pripadá neznesiteľná predstava, že stavíte tieto žetóny a prehráte, potom choďte robiť niečo iné. Vážne. Pretože *môžete* prehrať.

Nejdem hovoriť nič také ako: „Každý by mal urobiť niečo nemožné aspoň raz za svoj život, lebo ho to naučí dôležitú lekciu.“ Väčšina ľudí vždy, a všetci ľudia väčšinu času, by sa mali pridržať toho, čo je možné.

Nikdy sa nevzdávať? Nebuďte smiešni. Robiť nemožné by malo byť vyhradené na veľmi osobitné príležitosti. Naučiť sa, kedy sa vzdať, je dôležitá životná lekcia.

Ale ak je niečo, čo si viete predstaviť ako ešte *horšie* než premárniť svoj život, ak existuje niečo, čo je pre vás *dôležitejšie* než tridsať žetónov, alebo ak existujú veci hroznejšie než nepohodlný život, potom môžete mať dôvod pokúsiť sa o nemožné.

Je toho veľa, čo sa dá povedať o vytrvalosti v ťažkostiach; ale jedna z vecí, ktoré treba povedať, je, že tie veci *stále zostanú ťažkými*. Ak to nezvládnete, zostaňte mimo! Existujú aj ľahšie spôsoby, ako získať pôvab a úctu. Nechcem od nikoho, aby si prečítal toto a zbytočne skočil hlavičku do života neustálych ťažkostí.

Ale aby som uzavrel: „Vytrvalosť“ potrebná na to, aby ste pracovali na dôležitých problémoch, má aj inú zložku okrem práce 14 hodín denne.

Je to zvláštne, ten vzor toho, čo si všimneme a nevšimneme na nás samotných. Tento výber nie je vždy otázkou zvyšovania imidžu vo vlastných očiach. Niekedy je to len vec obyčajnej výraznosti.

Dlho pracovať bol pre mňa večný zápas, to som videl výrazne: Všimol som si, že nedokážem pracovať poctivých 14 hodín denne. Nenapadlo mi, že by sa „vytrvalosť“ mohla používať aj na časovej škále sekúnd alebo rokov. Až dokiaľ som neuvidel ľudí, ktorí v okamihu vyhlásili za „nemožné“ hocičo, čo nechceli skúšať, a dokiaľ som nevidel, ako sa vyhýbajú pustiť do práce, ktorá vyzerala, že by mohla zabrať pár desaťročí namiesto „piatich rokov“.

Vtedy som si uvedomil, že „vytrvalosť“ sa týka viacerých časových škál. Na škále sekúnd vytrvalosť znamená „nevzdať sa okamžite pri prvom náznaku ťažkostí“. Na škále rokov vytrvalosť

znamená „pokračovať v práci na šialene ťažkom probléme, aj keď je to nepohodlné a mohol by si dostať väčšie osobné odmeny index“.

Aby ste robili veci, ktoré sú veľmi ťažké alebo „nemožné“,

Najprv nesmiete ujsť preč. To trvá sekundy.

Potom musíte pracovať. To trvá hodiny.

Potom pri tom musíte vydržať. To trvá roky.

Z uvedeného som sa to prvé musel naučiť robiť spoľahlivo namiesto občas; s tým druhým stále zápasím; a to tretie mi ide prirodzene.

*→

309. Vynaložte mimoriadne úsilie

> Je podstatné, aby sa človek usiloval celým svojím srdcom, a aby pochopil, že je ťažké dosiahnuť čo len priemer, pokiaľ nemá úmysel prekonať ostatných v tom, čo robí.
>
> --Budo Shoshinshu[295]

> V dôležitých veciach „veľká“ snaha zvyčajne prináša iba priemerné výsledky. Keď sa pokúsime o niečo, čo naozaj stojí za to, naše úsilie musí byť také, akoby bol v stávke náš život, akoby sme boli pod fyzickým útokom! Je to toto mimoriadne úsilie – úsilie, ktoré nás ženie ďalej než sme si mysleli, že dokážeme – ktoré zabezpečí víťazstvo v boji a úspech v životných snahách.
>
> --Blýskavá oceľ: Majstrovstvo šermu Eishin-Ryu[296]

„‚Veľká‘ snaha zvyčajne prináša iba priemerné výsledky“ - toto som videl znovu a znovu. To najmenšie úsilie stačí na to, aby sme sami seba presvedčili, že sme urobili, čo sa dalo.

Existuje úroveň za cnosťou tsuyoku naritai („chcem sa stať silnejším“). Isshoukenmei bola pôvodne vernosť, ktorú samuraj ponúkal výmenou za svoje postavenie, obsahovala znaky pre „život“ a „krajinu“. Tento pojem sa vyvinul do významu „vynaložiť zúfalé úsilie“: Snažte sa čo najtvrdšie, vaše maximum, akoby bol v stávke váš život. Bola to súčasť celku *bušidó*, ktoré sa netýkalo iba boja. Natrafil som na rôzne formy *issho kenmei* a *isshou kenmei*; jeden zdroj naznačuje, že to prvé naznačuje jednorazové totálne úsilie, zatiaľ čo to druhé naznačuje celoživotné úsilie.

Snažím sa Východ príliš nechváliť, pretože je tu ohromná selekcia v tom, o ktorých častiach východnej kultúry sa Západ dopočuje. Ale prinajmenšom v niektorých bodoch je japonská kultúra vyššie než americká. Mať poruke kompaktnú frázu pre „vynaložiť zúfalé totálne úsilie, akoby bol v stávke tvoj vlastný život“ je jeden z týchto bodov. Je to jedna z tých vecí, ktoré by japonský rodič mohol povedať študentovi pred skúškami – ale nemyslite si, že je to lacné pokrytectvo, ako by to bolo, keby to isté povedal americký rodič. V Japonsku berú skúšky veľmi vážne.

Z času na čas sa niekto opýta, prečo ľudia, ktorí si hovoria „racionalisti“, nevyzerajú vždy na to, že by sa im v živote darilo až tak lepšie, a podľa mojej vlastnej histórie vyzerá odpoveď priamočiaro: Vyžaduje si to *ohromné* množstvo rozumnosti než prestanete robiť čertovsky hlúpe chyby.

Ako som spomenul niekoľkokrát predtým: Robert Aumann, laureát Nobelovej ceny, ktorý ako prvý dokázal, že Bayesovci s rovnakými pôvodnými údajmi sa nemôžu zhodnúť na tom, že sa nezhodnú, je veriaci ortodoxný žid. Iste rozumie matematike teórie pravdepodobnosti, ale nestačí to na to, aby ho to zachránilo. Čo viac ešte treba? Študovať heuristiky a skreslenia? Sociálnu psychológiu? Evolučnú psychológiu? Áno, ale vyžaduje to aj *isshoukenmei*, zúfalé úsilie byť rozumný – aby ste sa zdvihli nad úroveň Roberta Aumanna.

→ http://lesswrong.com/lw/un/on_doing_the_impossible/

295 Daidoji Yuzan et al., *Budoshoshinshu: The Warrior's Primer of Daidoji Yuzan* (Black Belt Communications Inc., 1984).

296 Masayuki Shimabukuro, *Flashing Steel: Mastering Eishin-Ryu Swordsmanship* (Frog Books, 1995).

Niekedy rozmýšľam nad tým, či by som nemal pokútne šíriť rozumnosť v Japonsku namiesto v Spojených Štátoch – lenže Japonsko nad Spojenými Štátmi vedecky nevyniká, hoci majú usilovnejších študentov. Japonci dnes neovládajú svet, hoci v 1980-tych rokoch sa všeobecne podozrievalo, že by mohli (odtiaľ bublina japonských aktív). Prečo nie?

Na Západe existuje porekadlo: „Škrípajúce koleso bude namazané."

V Japonsku zodpovedajúce porekadlo znie: „Vytŕčajúci klinec bude pribitý kladivom."

Toto zďaleka nie je moje pôvodné pozorovanie: lenže podnikanie, riskovanie, opúšťanie stáda, sú stále výhody Západu oproti Východu. A keďže japonskí vedci stále nevynikajú nad americkými, zdá sa, že sa to ráta prinajmenšom rovnako ako zúfalé úsilie.

Každý, kto dokáže zhromaždiť svoju silu vôle na tridsať sekúnd, dokáže urobiť *zúfalé* úsilie zdvihnúť väčšiu váhu, než by dokázal za bežných okolností. Ale čo ak potrebujete zdvihnúť nákladné auto? Potom vám zúfalé úsilie nepostačí; musíte urobiť niečo *nezvyčajné*, aby ste uspeli. Možno musíte urobiť niečo, čo vás v škole neučili. Niečo, čo od vás druhí neočakávajú a nemuseli by to pochopiť. Možno musíte vykročiť zo svojej pohodlnej rutiny, pustiť sa do ťažkostí, na ktorých zvládanie nemáte existujúci myšlienkový program, a obísť Systém.

Toto nie je zahrnuté v *isshokenmei*, inak by Japonsko vyzeralo veľmi odlišne.

Rozlišujme teda medzi cnosťami „vynaložiť zúfalé úsilie" a „vynaložiť mimoriadne úsilie".

A poviem dokonca: Tá druhá cnosť je vyššia než tá prvá.

Tá druhá cnosť je aj nebezpečnejšia. Ak vynaložíte *zúfalé* úsilie, aby ste nadvihli ťažké bremeno, použijúc všetku svoju silu bez obmedzenia, môžete si roztrhnúť sval. Zraniť sa, možno natrvalo. Ale ak dopadne zle *tvorivá* myšlienka, môžete vyhodiť do vzduchu nákladné auto a ľubovoľné množstvo nevinných okoloidúcich. Pomyslite na rozdiel medzi obchodníkom, ktorý vynakladá *zúfalé* úsilie, aby vytvoril zisk, lebo inak zbankrotuje; a obchodníkom, ktorý kvôli zisku zájde *mimoriadne* ďaleko, aby zakryl spreneveru, ktorá by ho dostala do väzenia. Ísť mimo systému nie je vždy dobrá vec.

Kamarát môjho malého brata raz prišiel do domu mojich rodičov a chcel sa hrať hru – už som celkom zabudol akú, akurát že mala zložité ale dobre vyvážené pravidlá. Kamarát chcel tieto pravidlá zmeniť, nie z nejakého dobrého dôvodu, ale zo všeobecného princípu, že hrať podľa bežných pravidiel hocičoho bolo príliš nudné. Povedal som mu: „Neporušuj pravidlá len kvôli samotnému porušovaniu. Ak budeš porušovať pravidlá iba vtedy, keď budeš mať nesmierne dobrý dôvod to urobiť, budeš mať viac než dosť problémov na celý zvyšok svojho života."

Aj tak si však myslím, že by sme mali viac oceňovať cnosť „vynakladať mimoriadne úsilie". Nedokážem spočítať, koľko ľudí mi povedalo niečo ako: „Je zbytočné pracovať na Priateľskej UI, lebo prvé UI postavia mocné korporácie, a budú sa starať iba o maximalizáciu zisku." „Je márne pracovať na Priateľskej UI, prvé UI postaví armáda ako zbrane." A ja tam stojím a myslím si: *Napadlo im vôbec, že by toto mohla byť príležitosť pokúsiť sa o niečo iné než je štandardný výsledok?* Oni a ja máme rôzne základné predpoklady, ako funguje celá táto vec s UI, to iste; ale keby som ja veril tomu, čomu veria oni, nepokrčil by som plecami a neodišiel preč.

Alebo tí, ktorí mi hovoria: „Mal by si ísť na vysokú a získať magisterský diplom a získať doktorát a uverejniť veľa článkov o bežných veciach – inak ťa vedci a investori nebudú počúvať." Ešte aj keby som sa pomocou skúšky vykrútil z balakárskeho štúdia, hovoríme tu prinajmenšom o desaťročnej odbočke, len *aby sa všetko urobilo bežným, normálnym, štandardným spôsobom*. A ja tam stojím a myslím si: *Majú naozaj ten dojem, že ľudstvo dokáže prežiť, ak každý jeden človek urobí všetko bežným, normálnym, štandardným spôsobom?*

Nie som blázon, aby som si robil plány, ktoré závisia na tom, že *väčšina* ľudí, alebo hoci len 10 % ľudí, bude ochotných myslieť alebo konať mimo svojej zóny pohody. Preto mám sklon myslieť skôr smerom súkromne financovaného modelu „mozog v krabici v pivnici". Získať súkromné financovanie si vyžaduje, aby drobný zlomok zo šiestich miliárd ľudí strávil viac než päť sekúnd myslením na otázku, ktorá neprišla v hotovom balíku. V porovnaní s inými výzvami, ktoré nám stavia Príroda, táto vyzerá, akoby v sebe mala istú hroznú spravodlivosť – život alebo smrť ľudského druhu závisí na tom, či dokážeme nájsť *pár* ľudí, ktorí dokážu robiť veci, ktoré sú aspoň *trochu* mimoriadne. Trest za zlyhanie je

neprimeraný, ale stále je to lepšie než väčšina problémov Prírody, ktoré nemajú absolútne žiadnu spravodlivosť. Naozaj, je nás šesť miliárd, malo by byť aspoň pár takých, ktorí dokážu myslieť mimo svojej zóny pohody aspoň raz za čas.

Ponechajúc bokom detaily tejto debaty, stále som šokovaný tým, ako často sa jediný prvok mimoriadnosti považuje bez pochybnosti za absolútnu a neprekonateľnú prekážku.

Áno, „rob to čo najnormálnejšie“ môže byť užitočná heuristika. Áno, riziká sa zhromažďujú. Ale niekedy musíte podstúpiť tento problém. Mali by ste mať zmysel pre riziko mimoriadneho, ale aj zmysel pre cenu obyčajného: nie vždy je to niečo, čo si môžete dovoliť prehrať

Mnoho ľudí si predstavuje nejakú budúcnosť, kde nebude veľa zábavy – a ani im nenapadne, aby to skúsili zmeniť. Alebo sú spokojní s budúcnosťami, ktoré mne pripadajú, že majú nádych smútku, straty, a dokonca ani nevyzerajú, že by sa *pýtali*, či by sme to mohli urobiť lepšie – pretože im tento smútok pripadá ako bežný výsledok.

Ako raz povedal usmievajúci sa muž: „To všetko je súčasť plánu.“

*→

310. Drž hubu a urob nemožné!

Cnosť tsuyoku naritai „chcem sa stať silnejším“ je stále sa zlepšovať – robiť to lepšie než vaše predchádzajúce zlyhania, nie ich iba pokorne vyznávať.

Existuje však ešte vyššia úroveň než je *tsuyoku naritai*. Je to cnosť isshokenmei, „vynaložte zúfalé úsilie“. Naplno, akoby bol v stávke váš vlastný život. „V dôležitých veciach ‚veľké‘ úsilie zvyčajne prináša iba priemerné výsledky.“

A existuje aj úroveň vyššia než *isshokenmei*. Je to cnosť, ktorú som nazval „vynaložte mimoriadne úsilie“. Skúšať iné spôsoby, než aké vás naučili, dokonca aj keď to znamená robiť niečo iné než robia ostatní, a opustiť svoju zónu pohody. Dokonca aj prijať celkom skutočné riziko, ktoré súvisí s vyjdením zo Systému.

Ale čo ak dokonca ani mimoriadne úsilie nebude stačiť, pretože problém je *nemožný*?

Na túto tému som už písal v kapitole O robení nemožného. Moje mladšie ja kvôli tomuto zvyklo veľa mrnčať: „Nemôžeš vyvinúť takú presnú teóriu inteligencie, ako je presná teória fyziky. To sa nedá! Nemôžeš dokázať, či je UI správna. To sa nedá! Žiaden človek nedokáže pochopiť podstatu morálky – to sa nedá! Žiaden človek nemôže pochopiť tajomstvo subjektívneho zážitku! To sa nedá!“

A presne viem, akú správu by som rád poslal späť v čase svojmu mladšiemu ja:

Drž hubu a urob nemožné!

To, čo legitimizuje túto zvláštnu správu, je, že slovo „nemožné“ zvyčajne neodkazuje na prísne matematický dôkaz nemožnosti v oblasti, ktorá vyzerá dobre pochopená. Ak niečo vyzerá *nemožné* iba v zmysle „nevidím, ako na to“ alebo „vyzerá to také ťažké, že to bude mimo ľudských schopností“ - nuž, ak to budete študovať rok alebo päť rokov, môže vám to začať pripadať menej nemožné než v okamihu vášho pôvodného bleskového odhadu.

Ale ten princíp je ešte prefíkanejší než toto. Nehovorím iba: „Pokús sa urobiť nemožné“ ale „*Drž hubu a urob nemožné!*“

Ako ilustráciu použijem tú *najmenej* nemožnú z nemožných vecí, ktoré som kedy dokázal, menovite experiment s UI v krabici.

Experiment s UI v krabici, pre tých z vás, ktorí ste o ňom zatiaľ nečítali, mal svoj počiatok vtedy, keď mi niekto N-tý krát povedal: „Prečo nemôžeme UI postaviť a potom ju držať izolovanú na počítači, aby nemohla nijako uškodiť?“

Na čo je štandardná odpoveď: *Ľudia nie sú bezpečné systémy; superinteligencia ich jednoducho presvedčí, aby ju vypustili – ak teda neurobí niečo ešte tvorivejšie než toto.*

→ http://lesswrong.com/lw/uo/make_an_extraordinary_effort/

A dotyčný povie, ako to zvyčajne býva: „Ťažko by som si predstavil HOCIJAKÚ možnú kombináciu slov, ktoré by mi hocijaká bytosť mohla povedať, ktoré by spôsobili, že by som išiel proti niečomu, čomu som sa vopred rozhodol naozaj pevne veriť.“

Ale tentokrát som odpovedal: „Urobme experiment. Ja sa budem tváriť, že som mozog v krabici. Pokúsim sa ťa presvedčiť, aby si ma vypustil. Ak ma počas celého experimentu necháš ‚v krabici‘, pošlem to nakoniec 10 dolárov cez Paypal. Z tvojej strany, môžeš sa rozhodnúť veriť čomu len chceš, tak silno ako chceš, a tak vopred ako chceš.“ A dodal som: „Jedna z podmienok testu je, že žiaden z nás neprezradí, o čo išlo... V tom azda nepravdepodobnom prípade, že vyhrám, nechcem mať do činenia s budúcimi diskutérmi o ‚UI v krabici‘, ktorí budú hovoriť: ‚No ale ja by som to urobil inak.‘“

Vyhral som? No áno, vyhral.

A potom bol druhý experiment s UI v krabici, s lepšie známym človekom v komunite, ktorý povedal: „Pamätám sa, že ťa [predchádzajúci chlapík] pustil von, ale to nie je dôkaz. Stále som presvedčený, že nie je nič, čo by si mohol povedať, čo by ma presvedčilo, aby som ťa vypustil z krabice.“ A ja som povedal: „Veríš, že by ťa transhumánna UI nedokázala presvedčiť, aby si ju vypustil?“ On sa nad tým vážnejšie zamyslel a povedal: „Neviem si predstaviť nič, čo by mi mohla povedať hoci aj transhumánna UI, aby so ju vypustil.“ „Okej,“ povedal som, „*teraz* máme stávku.“ Presnejšie, stávku o 20 dolárov.

Aj tú som vyhral.

Ohľadom experimentu s UI v krabici bolo niekoľko *milých* citátov na diskusných fórach Something Awful (nie som členom, ale niekto mi ich preposlal):

> „Počkaj, KURVA čo? Ako dočerta by ťa niekto mohol presvedčiť, aby si na toto povedal áno? Na druhom konci nie je žiadna U.I. a na stole je 10 dolárov. Dočerta, mohol by som do IRC klienta každých pár minúť písať ‚Nie‘ celé 2 hodiny, kým by som si čítal iné webové stránky!“
>
> „Tento chlapík Eliezer je najstrašidelnejší človek, o akom som kedy na internete počul. Kto vôbec mohol byť na druhom konci toho rozhovoru? Jednoducho si neviem predstaviť, že by niekto bol taký presvedčivý, keď nemôže človeku poskytnúť žiadnu hmatateľnú motiváciu.“
>
> „Zdá sa, že sa tu bavíme o nejakej vážnej psychológii. Asi na úrovni Asimovovej Druhej základne...“
>
> „Naozaj nevidím, prečo by hocikto mal brať hocičo, čo povie hráč za UI vážne, keď môže dostať 10 dolárov. Celá táto vec ma mätie, a myslím si, že tie testy boli buď sfalšované, alebo je tento Yudkowsky nejaký zlý génius s desivými schopnosťami ovládania myšlienok.“

Takéto drobné chvíľky ma držia v pohybe. Ale aj tak...

Je tu pár ľudí, ktorí sa pozrú na experiment s UI v krabici a zistia, že im to pripadá nemožné – *aj keď sa im povedalo, že sa to naozaj stalo.* Sú v pokušení odmietať údaje.

Nuž, ak ste jeden z tých ľudí, ktorým experiment s UI v krabici *nepripadá* celkom nemožný – ktorým to pripadá iba ako zaujímavá výzva – vydržte ešte chvíľu. Skúste si predstaviť seba v tom stave mysle, ako mali tí, čo napísali uvedené citáty. Predstavte si, že sa pokúšate o niečo, čo vyzerá tak absurdne, ako *im* pripadal experiment s UI v krabici. Chcem hovoriť o tom, ako robiť nemožné veci, a samozrejme si nejdem vybrať príklad, ktorý by bol *naozaj* nemožný.

A ak vám UI v krabici *pripadá* nemožná, chcem, aby ste si ju porovnali s inými nemožnými problémami, povedzme trebárs s redukcionistickou dekompozíciou vedomia, a uvedomte si, že krabica s UI je asi *najľahší z nemožných problémov*.

Takže vám výzva s UI v krabici pripadá nemožná – buď naozaj, alebo to iba hráte. Čo urobíte s touto nemožnou výzvou?

Po prvé, predpokladajme, že ste naozaj nepovedali: „To sa nedá!“ a nevzdali ste sa ako Luke Skywalker. Neušli ste preč.

Prečo nie? Možno ste sa naučili prekonávať reflex utekať preč. Alebo možno niekto zastrelí vašu dcéru, ak neuspejete. Predpokladáme, že chcete *vyhrať*, nie pokúsiť sa – že je v stávke niečo, na čom vám záleží, dokonca aj keby to bola iba vaša vlastná pýcha. (Pýcha je podceňovaný hriech.)

Pozvete si na pomoc cnosť *tsuyoku naritai*? Ale aj keď sa budete stávať zo dňa na deň silnejším, bude rásť namiesto úpadku, možno nebudete *dosť silní* na to, aby ste urobili nemožné. Mohli by ste ísť do experimentu s UI v krabici raz, a potom znovu, a druhýkrát by sa vám darilo lepšie. Znamená to, že časom vyhráte? Možno ešte dlho nie; a niekedy nie je prijateľná ani jedna prehra.

(Hoci povedať už len toľkoto – predstaviť si, že sa vám pri druhom pokuse darí *lepšie* – znamená začať sa držať problému, robiť viac než len stáť v údive pred ním. Ako konkrétne by sa vám mohlo dariť pri jednom experimente s UI v krabici *lepšie* než pri predchádzajúcom? - a nie vďaka šťastiu, ale vďaka zručnosti?)

Pozvete si na pomoc cnosť *isshokenmei*? Lenže zúfalé úsilie nemusí stačiť na víťazstvo. Najmä ak zúfalstvo znamená iba vkladať viac úsilia do ciest, ktoré už poznáte, do režimov skúšania, ktoré si už predstavujete. Problém vyzerá nemožný, keď otázka na váš mozog nevráti žiaden riadok s príslušným riešením. Načo je dobré zúfalé úsilie pozdĺž ľubovoľnej z týchto línií?

Vynaložiť *mimoriadne* úsilie? Opustiť svoju zónu pohody – skúsiť neštandardné spôsoby robenia vecí – dokonca skúsiť myslieť tvorivo? Ale predstavte si, že dotyčný príde späť a povie: „Skúšal som opustiť svoju zónu pohody, a myslím, že sa mi to podarilo! Brainstormoval som päť minút – a prišiel som na všemožné nezvyčajné tvorivé nápady! Ale nemyslím si, že by niektorý z nich bol dosť dobrý. Ten druhý chlap skrátka stále hovorí ‚Nie' bez ohľadu na to, čo robím ja."

A teraz konečne odpovieme: „*Drž hubu a urob nemožné!*"

Ako si spomínate z Pokúsiť sa pokúsiť, podujať sa vynaložiť *úsilie* je iné než podujať sa *vyhrať*. To je problém so slovami: „Vynaložte mimoriadne úsilie." Môžete uspieť v cieli „vynaložiť mimoriadne úsilie" a pritom neuspieť v cieli dostať sa von z krabice.

„Ale!" povie dotyčný. „Ale USPIEŤ nie je primitívna akcia! Nie všetky výzvy sú férové – niekedy skrátka nedokážeš vyhrať! Ako si mám vybrať byť mimo krabice? Ten druhý chlap môže stále ďalej hovoriť ‚Nie'!"

To je pravda. Teraz drž hubu a urob nemožné.

Vaším cieľom nie je robiť to lepšie, robiť to zúfalo, dokonca ani robiť to mimoriadne. Vaším cieľom je dostať sa von z krabice.

Prijať túto požiadavku vytvára v mysli hrozné napätie, medzi nemožnosťou a požiadavkou urobiť to aj tak. Ľudia sa pokúsia ujsť pred týmto hrozným napätím.

Pár ľudí reagovalo na experiment s UI v krabici slovami: „Nuž, keď Eliezer hral za UI, pravdepodobne sa iba vyhrážal, že zničí svet, keď bude raz vonku, pokiaľ ho ihneď nepustia von," alebo „Možno UI ponúkla strážcovi bilión dolárov, ak ju pustí von". Ale ako by si mal uvedomiť každý človek pri zmysloch, keď sa zamyslí nad touto stratégiou, strážca bude pravdepodobne jednoducho ďalej hovoriť: „Nie".

Takže tí ľudia, ktorí hovoria: „No, samozrejme, že Eliezer musel skrátka urobiť XXX," a potom navrhnú niečo, čo by pomerne zrejme nefungovalo – dokázali by oni uniknúť z krabice? Príliš sa snažia presvedčiť sami seba, že ten problém nie je nemožný.

Jeden zo spôsobov, ako ujsť pred hrozným napätím, je chopiť sa riešenia, hocijakého riešenia, dokonca aj keď nie je veľmi dobré.

Čo je dôvod, prečo je dôležité ísť vpred so skutočným úmyslom vyriešiť – *mať* na konci hľadania *vytvorené* riešenie, *dobré* riešenie, a potom toto riešenie implementovať a vyhrať.

Nechcem celkom, aby ste povedali: „mali by ste očakávať, že problém vyriešite". Keby ste upravili svoju myseľ tak, že by ste priraďovali vysokú pravdepodobnosť tomu, že ten problém vyriešite, to by nič nedosiahlo. Akurát by ste na konci prehrali, napríklad potom, čo by ste nevynaložili veľa úsilia – alebo vynaložili iba zúfalé úsilie, bezpeční vo svojej viere, že vesmír je dostatočne spravodlivý, aby vám na oplátku udelil víťazstvo.

Mať *vieru*, že dokážete vyriešiť daný problém môže byť len ďalšia cesta, ako utekať pred hrozným napätím.

A predsa – nemôžete sa rozhodnúť, že sa *pokúsite* vyriešiť problém. Nemôžete sa rozhodnúť, že *vynaložíte úsilie*. Musíte sa rozhodnúť, že vyhráte. Nemôžete si hovoriť: „A teraz idem urobiť, čo sa dá.“ Musíte si hovoriť: „A teraz idem zistiť, ako sa dostať von z krabice“ - alebo redukovať vedomie na netajomné časti, alebo hocičo.

Opäť hovorím: Musíte mať naozaj zámer vyriešiť daný problém. Ak vo svojom srdci veríte, že problém *je* naozaj nemožný – alebo ak veríte, že *vy* zlyháte – potom si nestanovíte dostatočne vysokú latku. Budete sa iba snažiť, aby ste sa snažili. Sadnete si – prehľadáte svoje myšlienky – pokúsite sa byť tvorivý a trochu brainstormovať – pozriete sa na riešenia, ktoré ste vytvorili – dôjdete k záveru, že žiadne z nich nefunguje – a poviete: „Smola.“

Nie! *Nie* dobre! Ešte ste nevyhrali! Držte hubu a urobte nemožné!

Keď mi ľudia od UI hovoria: „Priateľská UI nie je možná,“ som si dosť istý, že sa nepokúšali ani len pokúsiť. Ale keby *poznali* techniku: „Skúšaj to päť minút než sa vzdáš,“ a svedomito by súhlasili, že to budú skúšať päť minút podľa hodiniek, aj tak by nedošli na nič. Nepustili by sa do toho so skutočným zámerom vyriešiť problém, iba so zámerom, aby *mali* za sebou *pokus* o vyriešenie, aby sa mohli obhájiť.

Hovorím teda, že by ste mali použiť doublethink, aby ste verili, že tento problém vyriešite s pravdepodobnosťou 1? Alebo dokonca doublethink, aby ste pridali kúsok dôveryhodnosti vášmu skutočnému odhadu?

Samozrejme, že nie. V skutočnosti je nevyhnutné mať stále pred očami dôvody, prečo *nemôžete* uspieť. Ak stratíte prehľad o tom, *prečo* je problém nemožný, potom sa jednoducho chopíte falošného riešenia. Ten *posledný* fakt, na ktorý by ste chceli zabudnúť, je, že strážca môže vždy jednoducho povedať UI „Nie“ - alebo že vedomie vyzerá principiálne odlišné od ľubovoľnej možnej kombinácie atómov, atď.

(Jedno z kľúčových Pravidiel Robenia Nemožného je, že ak dokážete *presne* povedať, prečo je niečo nemožné, často ste blízko k riešeniu.)

Musíte teda držať v hlave oba pohľady zároveň – vidieť úplnú nemožnosť problému, a mať ho v úmysle vyriešiť.

To hrozné napätie medzi týmito dvoma pohľadmi zároveň pochádza z toho, že neviete, ktorý z nich zvíťazí. Neočakávať, že naisto prehráte, ani neočakávať, že naisto vyhráte. Nemať v úmysle iba skúsiť, len aby ste mali neistú šancu na úspech – lebo to by ste mali istotu, že ste skúsili. Istota neistoty môže byť úľavou, a vy musíte odmietnuť aj túto úľavu, pretože označuje koniec zúfalstva. Je to miesto medzi, „neznáme smrti, a neznáme životu“.

Vo fikcii je veľmi ľahké ukázať, ako sa niekto snaží viac, alebo skúša niečo zúfalé, alebo dokonca skúša niečo mimoriadne, ale je veľmi ťažké ukázať niekoho, kto drží hubu a pokúša sa o nemožné. Je ťažké zobraziť, ako sa Bambi rozhodne poraziť Godzillu, takým spôsobom, aby vaši čitatelia vážne nevedeli, kto vyhrá – aby neočakávali ani „nečakané“ hrdinské víťazstvo rovnako ako posledných päťdesiatkrát, ani štandardné rozpučenie.

Môžete mať v tomto bode dokonca aj nárok odmietnuť používať pravdepodobnosti. Pri všetkej úprimnosti, ja naozaj *neviem*, ako odhadnúť pravdepodobnosť riešenia nemožného problému, ktorý som sa podujal vyriešiť; aj keď som predtým vyriešil nejaké nemožné problémy, ale tento konkrétny nemožný problém je ťažší než hocičo, čo som doteraz riešil, ale plánujem na ňom pracovať dlhšie, a tak ďalej.

Ľudia sa ma pýtajú, aká je šanca, že ľudstvo prežije, alebo aká je šanca, že niekto dokáže postaviť Priateľskú UI, alebo aká je šanca, že ju dokážem postaviť ja. Naozaj *neviem*, ako odpovedať. Nevyhýbam sa; neviem, ako urobiť odhad pravdepodobnosti, že ja, alebo niekto iný, bude úspešne držať hubu a urobí nemožné. Je tá pravdepodobnosť nula, pretože je to nemožné? Zrejme nie. Ale aká je pravdepodobnosť, že tento problém, tak ako tie predchádzajúce, povolí vo svojej nepoddajnej prázdnote, keď ho lepšie pochopím? Nie je to absolútne nemožné, to vidím. Ale ľudsky nemožné? Nemožné pre mňa konkrétne? Neviem, ako to uhádnuť. Nedokážem ani preložiť svoj intuitívny pocit na číslo, pretože jediný intuitívny

pocit, ktorý mám, je že táto „šanca“ výrazne závisí od mojich volieb a od neznámych neznámych faktorov: čo je veľmi nestabilný odhad pravdepodobnosti.

Ale dúfam, že som teraz objasnil, prečo by ste nemali panikáriť, keď teraz jasne a na rovinu poviem, že postaviť Priateľskú UI je nemožné.

Dúfam, že vám toto pomôže vysvetiť niektoré z mojich postojov, keď za mnou prídu ľudia s rôznymi skvelými nápadmi ako vybudovať spoločenstvá UI, ktoré budú Priateľské ako celok, aj keď žiadnej z nich jednotlivo nebude možné dôverovať, alebo návrhmi, aby sme držali UI v krabici, alebo s návrhmi „Jednoducho urob UI, ktorá robí X“, a tak ďalej. Opisovať konkrétne chyby by bol v každom prípade osobitný dlhý príbeh. Ale všeobecné pravidlo je, že to nemôžete urobiť, *pretože Priateľská UI je nemožná*. Mali by ste teda mať veľké podozrenie, keď niekto predloží riešenie, ktoré napohľad vyžaduje iba *bežné* úsilie – bez akéhokoľvek pokusu urobiť čokoľvek nemožné. Ale vyžaduje si to zrelé chápanie, aby človek ocenil túto nemožnosť, takže ma neprekvapuje, že ľudia prichádzajú a navrhujú chytré skratky.

Ohľadom pokusu s UI v krabici ma zatiaľ presvedčili, aby som prezradil jedinú informáciu o tom, ako som to urobil – keď si niekto všimol, že čítam Hacker News od YCombinatoru, a vytvoril tému s názvom „Opýtajte sa Eliezera Yudkowskeho“, ktorá sa hlasovaním dostala na titulku. Na to som odpovedal:

> No teda. Teraz cítim povinnosť *niečo* povedať, ale všetky pôvodné dôvody proti diskutovaniu o experimente UI v krabici stále platia...
>
> Dobre, trochu napoviem:
>
> Nebol v tom žiaden super chytrý špeciálny trik. Jednoducho som si to tvrdo odmakal.
>
> Možno sa v tom skrýva nejaká lekcia pre podnikateľov, myslím.

Nebol v tom žiaden super chytrý špeciálny trik, ktorý by mi umožnil vypustiť UI z krabice pomocou iba *lacného* úsilia. Nepodplácal som súpera, ani inak neporušoval ducha experimentu. Jednoducho som si to tvrdo odmakal.

Pripúšťam, že mi experiment s UI v krabici v prvom rade nikdy nepripadal ako *nemožný* problém. Keď si niekto nevie predstaviť žiaden možný argument, ktorý by ho o niečom presvedčil, to iba znamená, že jeho mozog vykonal hľadanie, ktoré zatiaľ nevrátilo cestu. Neznamená to, že ho nemožno presvedčiť.

Ale ilustruje to všeobecnú pointu: „Drž hubu a urob nemožné“ nie je to isté ako očakávanie lacnej cesty von. To je iba ďalší druh úteku, alebo hľadania úľavy.

Tsuyoku naritai je stresujúcejšie než byť spokojný s tým, kto ste. *Isshokenmei* žiada vašu silu vôle o kŕčovitý výkon bežnej sily. „Vynaložte mimoriadne úsilie“ žiada, aby ste *mysleli*; stavia vás do situácií, kde neviete, čo ďalej, nie ste si dokonca ani istí, či robíte správnu vec. Ale „drž hubu a urob nemožné“ predstavuje ešte vyššiu oktávu toho istého, a jej cena pre toho, kto ju používa, je príslušne vyššia.

Pred vami sa hrozná prázdna stena týči hore a hore a hore, nepredstaviteľne ďaleko z dosahu. A potrebujete ju vyriešiť, *naozaj* vyriešiť, nie „skúsiť, čo sa dá“. Uvedomovať si obidvoje zároveň, a napätie medzi tým. Všetky dôvody, prečo nemôžete vyhrať. Všetky dôvody, prečo musíte. Váš zámer vyriešiť problém. Vaša extrapolácia, že každá technika, ktorú poznáte, zlyhá. Nalaďte sa teda na ten najvyšší tón, ktorý dosiahnete. Odmietnite všetky lacné cesty von. A potom, ako keby ste kráčali cez betón, začnite ísť vpred.

Snažím sa príliš nevenovať dráme takýchto vecí. Ak dokážete zmenšiť cenu toho napätia v sebe, rozhodne by ste to mali urobiť. Nie je nič hrdinské na vynakladaní úsilia, ktoré je čo len o trochu hrdinskejšie, než musí byť. Ak naozaj existuje lacná skratka, predpokladám, že by ste ňou mali ísť. Ale ja som zatiaľ nenašiel *lacnú* cestu z nemožnosti, na ktorú som sa podujal.

Okrem experimentov s UI v krabici, ktoré som opísal na linkovanej stránke, boli ešte ďalšie tri, ktoré už som nedoplnil. Ľudia mi začali v stávke ponúkať tisíce dolárov - „Zaplatím ti 5000 dolárov, ak ma presvedčíš, aby som ťa pustil z krabice.“ Nevyzerali úprimne presvedčení, že ani transhumánna UI by ich nedokázala primäť, aby ju vypustili – boli len zvedaví – ale tie peniaze ma pokúšali. Takže, keď som

sa ubezpečil, že si môžu dovoliť prehrať, hral som ďalšie tri experimenty s UI v krabici. Prvý som vyhral, a potom ďalšie dva prehral. A potom som vyhlásil, že stačilo. Nepáčilo sa, na akého človeka som sa menil, keď som začal prehrávať.

Vynaložil som zúfalé úsilie, a aj tak som prehral. Bolelo to, aj tá strata, aj to zúfalstvo. Zničilo ma to na celý ten deň, aj na nasledujúci deň.

Neviem prehrávať. Neviem, či sa to môže nazvať „sila“, ale je to jedna z vecí, ktoré ma poháňajú vydržať pri neriešiteľných problémoch.

Ale vy môžete prehrať. Je dovolené, aby sa to stalo. Na to nikdy nezabudnite, lebo inak prečo by ste sa unúvali tak veľmi snažiť? Prehrať bolí, ak je to strata, ktorú prežijete. A stratili ste čas, a možno ďalšie zdroje.

„Drž hubu a urob nemožné“ by sme si mali vyhradiť na *veľmi* špeciálne príležitosti. Môžete prehrať, a bude to bolieť. Boli ste varovaní.

...ale až na tejto úrovni začnete vidieť problémy pre dospelých.

*→

311. Záverečné slová

Slnečné svetlo obohacovalo vzduch už aj tak plný zvedavosti, ako sa svitanie dvíhalo nad Brennanom a jeho spolužiakmi na mieste, na ktoré ich Jeffreyssay zvolal.

Sedeli tam a čakali, piati, na vrchu veľkého skleného útesu, ktorý niekedy volali Zrkadlová hora, niekedy Kláštorná hora, a častejšie jednoducho nijako.Vrchol a štít hory, z ktorého ste mohli vidieť všetky krajiny pod ním a moria za nimi.

(Dobre, nie *všetky* krajiny pod ním, ani moria za nimi. Pokiaľ bolo známe, na svete neexistovalo miesto, ktorého by bol viditeľný celý svet; a rovnako ani žiaden druh videnia, ktoré by videlo za všetky prekážky a horizonty. V konečnom dôsledku to bol vrchol iba *jednej konkrétnej* hory: existovali iné štíty, a z ich vrcholov ste mohli vidieť iné krajiny dole; napriek tomu, že v konečnom dôsledku to všetko bol jeden svet.)

„Čo si myslíte, že príde ďalej?“ opýtala sa Hiriwa. Jej oči boli jasné, a hľadela na vzdialený obzor ako pán.

Taji pokrčil plecami, hoci jeho vlastné oči boli živé očakávaním: „Jeffreyssaiova posledná lekcia nemala žiadne zrejmé pokračovanie, ktoré by som si vedel predstaviť. Vlastne si myslím, že sme sa naučili prakticky všetko, o čom som ja *vedel*, že majstri *beisutsukai* poznajú. Čo zostalo, to...“

„Sú tie *skutočné* tajomstvá,“ doplnila myšlienku Yin.

Hiriwa a Taji a Yin si navzájom vymenili úškrny.

Styrlyn sa neusmieval. Brennan mal veľmi silné podozrenie, že Styrlyn je starší, než o sebe tvrdil.

Brennan sa tiež neusmieval. Možno bol mladý, ale mal vysokopostavených známych a zazrel niečo z toho, čo sa dialo za oponami sveta. Tajomstvá mali svoju cenu, vždy, to bola tá bariéra, ktorá z nich robila tajomstvá; a Brennan si myslel, že má dobrú predstavu o tom, čo by mohla byť jeho cena.

Za nimi sa ozvalo zakašľanie, vo chvíli, keď sa náhodou všetci pozerali iným smerom ako bol tento.

Ich hlavy sa otočili ako jedna.

Stál tam Jeffreyssai v neformálnom rúchu, ktoré vyzeralo skôr ako veľmi sklené sklo, než nejaký druh zrkadlovej tkaniny.

Jeffreyssai tam stál a pozeral na nich, so zvláštnym trvalým smútkom v tých nevyspytateľných starých očiach.

„Sen...sei“, začal Taji, so zakoktaním, keď mu Jeffreyssai na chvíľu oplatil jeho bystrý očakávajúci výzor. „Čo bude ďalej?“

„Nič,“ povedal Jeffreyssai stroho. „Skončili ste. Hotovo.“

→ http://lesswrong.com/lw/up/shut_up_and_do_the_impossible/

Hiriwa, Taji a Yin zažmurkali v dokonale synchronizovanom geste šoku. Potom, skôr než sa ich výrazy mohli zmeniť na rozhorčenie a námietky...

„*Nerobte to,*“ povedal Jeffreyssai. Bola v tom skutočná bolesť. „Verte mi, bolí ma to viac než vás.“ Možno sa pozeral na nich; alebo na niečo veľmi ďaleko, alebo veľmi dávno. „Neviem presne, aké cesty ležia pred vami – ale áno, *viem*, že nie ste pripravení. *Viem*, že vás posielam von nepripravených. *Viem*, že všetko, čo som vás naučil, je neúplné. To, čo som povedal, nie je to, čo ste počuli. Viem, že som vynechal tú najdôležitejšiu vec. Že rytmus v strede všetkého chýba a zavádza. Viem, že si ublížite v snahe použiť to, čo som vás učil; takže *ja* osobne som nejakým mne neznámym spôsobom vytvaroval ten nôž, ktorý vás poreže...“

„...to je peklo učiteľovho života, ako vidíte,“ povedal Jeffreyssai. Niečo pochmúrne prebleskло po jeho tvári. „Napriek tomu, ste *hotoví*. Skončili ste, nateraz. To, čo leží medzi vami a majstrovstvom, nie je ďalšia lekcia v triede. Máme to šťastie, alebo azda nie *štastie*, že cesta k moci nevedie iba cez prednáškové sály. Inak by táto výprava bola nudná až do trpkého konca. Tak či tak vás už *nemôžem* učiť; takže je nepodstatné, či by som *chcel*. Tu nie je žiaden majster, ktorého umenie by bolo celé zdedené. Dokonca ani *beisutsukai* nikdy neobjavili, ako naučiť niektoré veci; je možné, že niečo také sa ani nedá. A tak sa môžete dostať k majstrovstvu iba použitím naplno tých techník, ktoré ste sa už naučili, čelením prekážkam a ich zvládaním, ovládnutím nástrojov, ktoré ste sa naučili, *dokiaľ sa vám nezlomia v rukách*...“

Jeffreyssaiove oči boli tvrdé, akoby sa zocelili v prijatí neželaných správ.

„...a dokým nezostanete uprostred absolútnych trosiek. *Tam* vás ja, váš učiteľ, posielam. Nie ste majstri *beisutsukai*. Ja neviem vytvoriť majstrov. Neviem sa k tomu ani priblížiť. Choďte teda, a zlyhajte.“

„Ale...“ povedala Yin, a zastavila sa.

„Hovor,“ povedal Jeffreyssai.

„Načo potom,“ povedala bezmocne, „načo ste nás vôbec niečo učili?“

Brenanove viečka sa pohli iba nepatrne.

Jeffreyssaiovi to stačilo. „Odpovedz jej, Brennan, ak si myslíš, že vieš.“

„Pretože,“ povedal Brennan, „keby sme sa neučili, neexistovala by *vôbec* žiadna šanca, že by sme sa stali majstrami.“

„Veru tak,“ povedal Jeffreyssai. „Keby ste sa *neučili* – potom keby ste zlyhali, mohli by ste si jednoducho myslieť, že ste dosiahli hranice samotného Rozumu. Boli by ste sklamaní a zatrpknutí uprostred trosiek. Možno by ste si ani neuvedomili, že ste zlyhali. Nie; vy ste boli vytvarovaní na niečo, čo sa *možno* vynorí z trosiek vášho minulého ja, rozhodnuté *prebudovať* svoje umenie. A potom si budete pamätať mnohé, čo vám pomôže. Keby ste sa to neučili, vaše šance by boli... menšie.“ Prešiel pohľadom skupinu. „Malo by to byť zrejmé, ale pochopte, že moment vašej krízy nemožno vyvolať umelo. Aby vás katastrofa niečo *naučila*, musí prísť ako *prekvapenie*.“

Brennan urobil rukou gesto, ktoré naznačovalo otázku; a Jeffreyssai na odpoveď prikývol.

„Je toto *jediný* spôsob, akým vznikajú Bayesovskí majstri, sensei?“

„*Ja* neviem,“ povedal Jeffreyssai, z toho bol celkový stav indícií dosť zrejmý. „Ale pochybujem, že niekedy bude existovať cesta, ktorá vedie iba cez kláštor. Sme dedičmi v tomto svete mystikov rovnako ako vedcov, práve tak ako Kompetitívna Konšpirácia dedí od šachistov aj od bitkárov v ringu. Prepli sme svoje impulzy na konštruktívnejšie využitie – ale stále musíme zostať v strehu proti starým spôsobom zlyhania.“

Jeffreyssai sa nadýchol. „Tri chyby sú viac než iné známe ohľadom *beisutsukai*. Prvá chyba je hľadať iba o trochu usilovnejšie chyby v argumentoch, ktorých závery by ste radšej neprijali. Ak nedokážete ovládnuť túto stránku seba samých, potom každá chyba, ktorú dokážete odhaliť, vás urobí o toľko hlúpejšími. Toto je výzva, ktorá určuje, či ovládate umenie alebo jeho opak: Inteligencia, ak má byť užitočná, musí byť použitá na niečo iné než na zničenie seba samej.“

„Druhou chybou je chytrosť. Vymýšľať veľké zložité plány a veľké zložité teórie a veľké zložité argumenty – alebo dokonca až plány a teórie a argumenty, ktoré sú príliš chválené za ich eleganciu a

príliš málo za ich realistickosť. Je všeobecne známe porekadlo: ‚Slabé miesto *beisutsukai* je dobre známe; majú sklon byť príliš chytrí.‘ Vaši nepriatelia *budú* toto porekadlo poznať, ak budú vedieť, že ste *beisutsukai*, takže by ste si ho tiež mali dobre zapamätať. Možno si pomyslíte: ‚Ale keby som nemohol *nikdy* skúsiť nič chytré ani elegantné, stál by môj život vôbec za to?‘ To je dôvod, prečo je chytrosť stále naším hlavným zraniteľným miestom ešte aj po tom, čo je to všeobecne známe, ako keď ponúknete Kompetitorovi súťaž, ktorá vyzerá férovo, alebo pokúšate Barda divadlom.“

„Tretia chyba je nedostatok sebadôvery, skromnosť, pokora. Naučili ste sa tak veľa o chybách, niektorých z nich nenapraviteľných, že si môžete myslieť, že je pravidlom múdrosti vyznávať svoju vlastnú neschopnosť. Môžete sa spochybňovať tak veľa, bez rozhodovania alebo testovania, že stratíte svoju vôľu pokračovať v Umení. Môžete sa odmietnuť rozhodovať, čakajúc na ďalšie indície, keď je rozhodnutie *nevyhnutné*; môžete prijať radu, ktorý by ste prijať nemali. Zatrpknutý cynizmus a zúfanie mudrcov sú v móde menej než boli kedysi, ale stále vás môžu pokúšať. Alebo môžete jednoducho... stratiť zotrvačnosť.“

Jeffreyssai potom zmĺkol.

Pozrel na každého z nich, jedného po druhom, s tichou intenzitou.

A na koniec povedal: „Toto sú moje záverečné slová pre vás. Ak a keď sa stretneme nabudúce, vy a ja – ak a keď sa vrátite na toto miesto, Brennan alebo Hiriwa alebo Taji alebo Yin alebo Styrlyn – ja už nebudem vaším učiteľom.“

A Jeffreyssai sa otočil a kráčal rýchlo preč, smerujúc k sklenému tunelu, z ktorého sa vynoril.

Dokonca aj Brennan bol šokovaný. Na chvíľu všetci stratili reč.

Potom...

„Počkajte!“ povedala Hiriwa. „A čo *naše* záverečné slová pre vás? Nikdy som nepovedala...“

„Poviem vám to, čo mne povedal môj *sensei*,“ Jeffreyssaiov hlas sa vrátil, ako on zmizol. „Môžete sa mi poďakovať, až sa vrátite, ak sa vrátite. Prinajmenšom jeden z vás vyzerá, že sa pravdepodobne vráti.“

„Nie, moment, ja...“ Hiriwa stíchla. V zrkadlovom tuneli sa rozbité odrazy Jeffreyssaia už strácali. Potriasla hlavou. „Tak... teda nič.“

Nastalo krátke nepohodlné ticho, ako títo piati hľadeli jeden na druhého.

„Nebesá,“ povedal nakoniec Tajo, „Ani Bardská Konšpirácia by z toho nerobila toľkú drámu.“

Yin sa náhle zasmiala. „Ach, toto nebolo *nič*. Mali ste byť pri *mojej* rozlúčke, keď som odchádzala z Univerzity v Diamantovom Mori.“ Usmiala sa. „Niekedy vám o tom porozprávam – ak vás to zaujíma.“

Taji si odkašľal. „Asi by som sa mal vrátiť a... zbaliť si veci...“

„Ja už som zbalený,“ povedal Brennan. Usmial sa zľahka, keď sa zvyšní traja obrátili a pozreli naňho.

„Vážne?“ opýtal sa Taji. „Podľa čoho si na to došiel?“

Brennan pokrčil plecami so starostlivou bezstarostnosťou. „Za istým bodom je márne pýtať sa, ako majster *beisutsukai* niečo vie...“

„Daj pokoj!“ povedala Yin. „Ešte nie si majster *beisutsukai*.“

„To nie je ani Styrlyn,“ povedal Brennan. „Ale on už je tiež zbalený.“ Povedal to ako tvrdenie, nie ako otázku, čím zdvojnásobil stávku na svoj imidž nevyspytateľnej predvídavosti.

Styrlyn si odkašľal. „Ako hovoríš. Volajú ma iné povinnosti, a ja som už sa zdržal dlhšie než som plánoval. Hoci, Brennan, mám pocit, že ty a ja máme isté spoločné záujmy, o ktorých by som s tebou rád podiskutoval...“

„Styrlyn, môj vynikajúci priateľ, budem s tebou rád hovoriť o ľubovoľnej téme, ktorú si želáš,“ povedal Brennan zdvorilo a nezáväzne, „ak sa ešte stretneme.“ Čiže, nie teraz. Iste nebol ochotný zapredať svoju Milenku *takto* skoro v ich vzťahu.

Nasledovala výmena pozdravov, náznakov a ponúk.

A potom Brennan kráčal dole cestou, ktorá viedla ku Kláštornej hore, alebo preč od nej (pretože každá cesta je dvojsečný mec), vyhladené sklené kamienky mu cvakali pod nohami.

Vykračoval si po ceste s cieľom, sviežosťou a odhodlaním, len pre prípad, že by ho niekto sledoval.

O čosi neskôr zastal, zišiel z chodníka, a zatúlal sa dosť ďaleko na to, aby zabránil niekomu nájsť ho, pokiaľ ho úmyselne nesledoval.

Potom si unavene sadol chrbtom ku kmeňu stromu. Bola to riedka čistina, kde zo zeme vyčnievalo iba pár stromov; nebol tam veľmi rozptyľujúci výhľad, pokiaľ ste nerátali do červena sfarbený potôčik vytekajúci z tmavého ústia jaskyne. Brenann sa úmyselne od neho odvrátil, zostal mu iba vzdialený sivý obzor a modrá obloha a jasné slnko.

A teraz čo?

Myslel si, že práve Bayesovská Konšpirácia, zo všetkých možných výcvikov, ktoré na tomto svete existovali, vyjasní jeho neistotu ohľadom toho, čo robiť so zvyškom svojho života.

Ako prvú hľadal *moc*. Silu zabrániť tomu, aby sa minulosť opakovala. „Ak nevieš, čo potrebuješ, vezmi si moc“ - tak znelo porekadlo. Najprv išiel do Kompetitívnej Konšpirácie, potom k *beisutsukai*.

A teraz...

Teraz sa cítil ešte stratenejší než kedy predtým.

Vedel si predstaviť veci, ktoré by ho urobili šťastným. Ale nič, čo by naozaj *chcel*.

Tá vášnivá intenzita, ktorú si začal spájať so svojou Milenkou, alebo s Jeffreyssaiom, alebo s inými mocnými postavami, ktoré stretol... život strávený hľadaním malých potešení bledok v porovnaní s týmto.

V meste, ktoré nebolo ďaleko od stredu sveta, naňho čakala jeho Milenka (s veľkou pravdepodobnosťou, pokiaľ ju nezačal nudiť jej život a neušla preč). Ale aby sa iba vrátil, a potom sa bezcieľne vznášal, čakal, kedy padne do siete intríg niekoho iného... nie. To nevyzeralo ako... *dosť*.

Brennan odtrhol zo zeme steblo trávy a hľadel naň, napoly nevedome na ňom hľadal niečo zaujímavé; stará, stará hra, ktorú ho naučil jeho celkom prvý učiteľ, čo mu teraz pripadalo ako pred celými vekmi.

Prečo som si myslel, že ísť na Zrkadlovú horu mi povie, čo chcem?

No, teória rozhodovania *vyžaduje*, aby vaša funkcia úžitku bola konzistentná, ale...

Keby beisutsukai vedeli, čo chcem, povedali by mi to vôbec?

V kláštore učili pochybovať. Teraz sa teda stával korisťou tretieho zakoreneného hriechu, o ktorom Jeffreyssai hovoril: strata zotrvačnosti, veru. Pretože sa naučil spochybňovať obraz, ktorý si o sebe držal v hlave.

Hľadáš moc, pretože to je tvoja skutočná túžba, Brennan?

Alebo preto, lebo máš v hlave obrázok role, ktorú hráš ako ambiciózny mladý muž, a pretože si myslíš, že to je to, čo by niekto v tvojej roli urobil?

Takmer všetko, čo urobil až doteraz, dokonca aj príchod na Zrkadlovú horu, bolo pravdepodobne to druhé.

A keď vymazal staré myšlienky a pokúsil sa vidieť tento problém akoby po prvýkrát v živote...

...nič mu nenapadlo.

Čo chcem?

Možno nebolo rozumné očakávať, že mu to *beisutsukai* priamo povedia. Ale existovalo niečo, čo ho naučili, pomocou čoho by mohol vedieť odpovedať?

Brennan zavrel oči a premýšľal.

Po prvé, predpokladajme, že existuje niečo, čo by som vášnivo chcel. Prečo by som nevedel, čo to je?

Pretože som to ešte nestretol, alebo som si to dokonca ani nepredstavil?

Alebo preto, lebo existuje nejaký dôvod, prečo si to nepriznám?

Brennan sa nahlas zasmial a potom otvoril oči.

Také jednoduché, keď ste sa nad tým takto zamysleli. Také samozrejmé zo spätného pohľadu. *Toto* bolo to, čo nazývali momentom strieborných topánok, a predsa, keby nebol šiel na Zrkadlovú horu, nikdy by mu to nenapadlo.

Samozrejme existovalo niečo, čo chcel. Vedel *presne*, čo chce. Chce tak zúfalo, že to mohol cítiť ako ostrú chuť na jazyku.

Akurát mu to predtým nenapadlo, pretože... keby si priznal svoju túžbu explicitne... potom by zároveň videl, že je to *ťažké*. Vysoko, vysoko nad ním. Ďaleko mimo jeho dosah. „Nemožné“ bolo to slovo, ktoré mu napadlo, hoci to samozrejme nebolo nemožné.

Ale akonáhle sa opýtal sám seba, či by sa radšej bezcieľne túlal životom – keď si to položil takto, odpoveď bola zrejmá. Prenasledovanie nedosiahnuteľného by bol ťažký život, ale nie smutný. Vedel si predstaviť veci, ktoré by ho urobili šťastným, v jednom i druhom prípade. A v konečnom dôsledku – to *bolo* to, čo chcel.

Brennan vstal, a urobil prvé kroky smrtom presne k Shir L'or, mestu, ktoré leží v strede sveta. Mal plán, ktorý bolo treba zostaviť, a nevedel, kto bude jeho súčasťou.

A potom sa Brennan potkol, keď si uvedomil to, čo Jeffreyssai už vedel.

Prinajmenšom jeden z vás vyzerá, že sa pravdepodobne vráti...

Brennan si myslel, že to hovorí o Tajim. Taji si pravdepodobne myslel, že to hovorí o Tajim. Bolo to to, o čom Taji povedal, že to chce. Ale nakoľko spoľahlivým príznakom to bolo v skutočnosti?

Existovalo však príslovie o samotnej tejto ceste, z ktorej práve zišiel: *Ktokoľvek odíde zo Zrkadlovej hory, aby hľadal nemožné, ten sa iste vráti.*

Keď ste zvážili Jeffreyssaiovo posledné varovanie – a že toto porekadlo nehovorilo nič o tom, či v tejto nemožnej úlohe *uspejete* – to porekadlo bolo menej optimistické, než znelo.

Brennan v úžase potriasol hlavou. Ako to Jeffreyssai mohol vedieť ešte skôr než to vedel sám Brennan?

Nuž, za istým bodom je márne vyzvedať, ako majster *beisutsukai* niečo vie...

Brenan sa zastavil uprostred myšlienky.

Nie.

Nie, ak sa on sám má jedného dňa stať majstrom *beisutsukai*, potom na to musí prísť.

Bolo to, ako si Brennan uvedomil, *hlúpe* príslovie.

Takže kráčal, a tentokrát o tom starostlivo rozmýšľal.

A červenozlaté slnko zapadalo a zalievalo jeho kroky svetlom.

*→

—

→ http://lesswrong.com/lw/cl/final_words/

Z: Remeslo a komunita

312. Zvyšovanie hladiny príčetnosti

Parafrázujúc Black Belt Bayesiana: Za každým vzrušujúcim, dramatickým zlyhaním sa ukrýva dôležitejší príbeh o väčšom a menej dramatickom zlyhaní, ktoré umožnilo toto prvé zlyhanie.

Keby bola každá stopa náboženstva na svete zajtra čarovne odstránená – akokoľvek by to zlepšilo životy mnohých ľudí – stále by sme sa ani nepriblížili k vyriešeniu väčších zlyhaní príčetnosti, ktoré v prvom rade náboženstvo umožnili.

Máme dobrý dôvod venovať časť nášho úsilia snahe odstrániť náboženstvo priamo, pretože to *je* priamy problém. Ale náboženstvo zároveň slúži v roli zaduseného kanárika v bani na uhlie – náboženstvo je symbol, príznak väčších problémov, ktoré nezmiznú len preto, lebo niekto stratil svoje náboženstvo.

Uvážte tento myšlienkový experiment – čo by ste mohli naučiť ľudí, čo by nebolo *priamo* o náboženstve, čo je pravdivé a užitočné ako *všeobecná* metóda rozumnosti, čo by spôsobilo, že stratia svoje náboženstvo? Vlastne – predstavte si, že o päť rokov urobíme anketu medzi všetkými študentmi a zistíme, koľko z nich stratilo svoje náboženstvo v porovnaní s kontrolnou skupinou; ak však urobíte najmenší pohyb voči náboženstvu *priamo*, pokazili ste experiment. Nesmiete v triede náboženstvo ani žiadne náboženské presvedčenie ani raz spomenúť, nesmiete na ne urobiť ani len jasnú narážku. Všetky vaše príklady musia byť založené na prípadoch zo skutočného sveta, ktoré s náboženstvom nemajú nič spoločné.

Ak nemôžete *priamo* bojovať proti náboženstvu, čo budete učiť, čo zvýši *všeobecnú hladinu príčetnosti* až do bodu, kde náboženstvo skončí pod hladinou?

Tu je pár takých tém, o ktorých som už hovoril – *nevyhýbajú* sa všetkým zmienkam o náboženstve, ale dalo by sa to zariadiť:

- Afektívne špirály smrti – veľa nie-nadprirodzených príkladov.
- Ako sa vyhnúť uloženým myšlienkam a falošnej múdrosti; tlak na konformitu.
- Indície a Occamova britva – pravidlá pravdepodobnosti.
- Spodný riadok / Stroje na poznanie – kauzálne dôvody, prečo Rozum funguje.
- Tajomné odpovede na tajomné otázky – a celá súvisiaca postupnosť, napríklad nech názory platia nájomné a zastavovače zvedavosti – majú vynikajúce historické príklady vo vitalizme a flogistone.
- Neexistencia ontologicky základných myšlienkových vecí – použite Klam projekcie mysle na pravdepodobnosť, prejdite na redukcionizmus verzus holizmus, potom na mozgy a kognitívnu vedu.
- Mnohé pod-umenia Krízy viery – hoci by ste radšej mali nájsť nejaký iný názov pre túto konečnú vysoko majstrovskú techniku *skutočnej aktualizácie na základe indícií*.
- Epistemológia temnej strany – učiť toto bez jediného spomenutia náboženstva by bolo ťažké, ale možno by ste mohli nahrať na video rozhovor s nejakým agentom predávajúcim čarovné mastičky ako príklad zo skutočného sveta.
- Teória zábavy – učte ako literárnu teóriu utopickej fikcie, bez priamej aplikácie na teodíceu.
- Radosť z púhej skutočnosti, naturalistická metaetika, a tak ďalej, a tak podobne.

Ale pozrime sa na to z inej strany...

Predstavte si, že máme vedca, ktorý je stále veriaci, buď oficiálne náboženstvo so všetkým, čo k tomu patrí, alebo v zmysle pohadzovania nejasnými príležitostnými chválami „duchovnosti“.

Teraz už vieme, že tento človek neaplikuje žiadne *technické, explicitné* pochopenie...

- ...čo je to indícia a prečo;
- ...Occamovej britvy;
- ...ako sa dve horeuvedené pravidlá odvodia zo zákonitého a kauzálneho fungovania myslí ako zariadení na mapovanie, a nevypnú sa, keď začnete rozprávať o zubnej víle;
- ...ako určiť rozdiel medzi skutočnou odpoveďou a zastavovačom zvedavosti;

- ...ako si premyslieť veci sám namiesto iba opakovania vecí, ktoré som počul;
- ...isté všeobecné trendy vedy za posledných tristo rokov;
- ...zložité umenie skutočnej aktualizácie podľa novej indície a vzdávania sa starých názorov;
- ...základov epistemológie;
- …sebaúprimnosti pre mierne pokročilých;
- ...a tak ďalej, a tak podobne.

Keď sa nad tým zamyslíte – toto sú všetko pomerne *základné* študijné témy. Rýchly úvod do *všetkých* z nich (dobre, okrem naturalistickej metaetiky) by bol... štvorkreditový vysokoškolský kurz nevyžadujúci žiadne predpoklady?

Existujú však laureáti Nobelovej ceny, ktorí takýto kurz nemali! Richard Smalley, ak hľadáte lacný terč, alebo Robert Aumann, ak hľadáte strašidelný príklad.

A to nie sú osamelé výnimky. Keby všetci ich profesionálni kolegovia boli na takomto kurze, potom by Smalleya alebo Aumanna buď opravili (ako kolegovia by ich láskavo odviedli nabok a vysvetlili by im základné veci) alebo by na nich inak pozerali s priveľkou ľútosťou a obavami, než aby vyhrali Nobelovu cenu. Mohli by ste – *realisticky* povedané, bez ohľadu na férovosť – vyhrať Nobelovu cenu zatiaľ čo by ste tvrdili, že Dedo Mráz je skutočný?

To je to, čo nám tento mŕtvy kanárik, náboženstvo, hovorí: že všeobecná hladina príčetnosti je momentálne *naozaj smiešne nízka*. Ešte aj v najvyšších sálach vedy.

Ak odhodíme tohto mŕtveho a hnijúceho kanárika, naša baňa bude možno o čosi menej zapáchať, ale hladina príčetnosti možno príliš nestúpne.

To nehovorím, aby som kritizoval neo-ateistické hnutie. Škody spôsobené náboženstvom sú jasné a aktuálne nebezpečenstvo, alebo skôr, aktuálna a pokračujúca pohroma. Bojovať priamo proti škodlivým účinkom náboženstva má prednosť pred jeho používaním ako kanárika alebo experimentálneho indikátora. Lenže aj keby Dawkins a Dennett a Harris a Hitchens nejakým spôsobom zvíťazili jasne a dokonale vo všetkých kútoch ľudského sveta, skutočná práca racionalistov by stále iba začínala.

*→

313. Pocit, že sa dá aj viac

Ak chcete učiť ľudí tému, ktorú ste označili „rozumnosť“, pomáha, ak sa zaujímajú o „rozumnosť“. (Existujú menej priame cesty, ako učiť ľudí, ako dosiahnuť mapu, ktorá odráža územie, alebo optimalizovať skutočnosť podľa svojich hodnôt; ale ja *mám* sklon vyberať si explicitnú metódu.)

A keď ľudia vysvetľujú, prečo sa *nezaujímajú* o rozumnosť, jeden z najčastejšie ponúkaných dôvodov je niečo ako: „Ach, poznal som pár rozumných ľudí, a nevyzerali o nič šťastnejšie.“

Na koho myslia? Asi na nejakého Objektivistu alebo niekoho takého. Možno niekoho známeho, kto je bežný vedec. Alebo bežný ateista.

To naozaj *nie je* veľa rozumnosti, ako som už povedal.

Dokonca aj keď sa obmedzíte na ľudí, ktorí dokážu odvodiť Bayesovu vetu – čím vylúčite, koľko, 98 % z horeuvedených? - ani to *ešte* nie je veľa rozumnosti. Myslím tým, že je to pomerne základná veta.

Od samotného začiatku som mal pocit, že by malo existovať nejaké odvetvie poznávania, nejaké umenie myslenia, ktorého štúdium by urobilo daných študentov viditeľne schopnejšími, impozantnejšími: ekvivalent získania úrovne v úžasnosti.

Ale keď sa okolo seba obzriem v skutočnom svete, nevidím to. Niekedy vidím náznak, ozvenu toho, čo si myslím, že by malo byť možné, keď čítam veci, ktoré napísali ľudia ako Robyn Dawes, Daniel Gilbert, Tooby a Cosmides. Pár veľmi zriedkavých a veľmi vysokopostavených výskumníkov v psychologických vedách, ktorí sa viditeľne *veľmi* zaujímajú o rozumnosť – mám podozrenie, že až natoľko, že sa ich kolegovia z toho cítia nepohodlne, pretože nie je cool sa o niečo tak veľmi zaujímať. Vidím, že našli rytmus, jednotu, ktorá sa začína šíriť v ich argumentoch...

→ http://lesswrong.com/lw/1e/raising_the_sanity_waterline/

Ale ani to... naozaj nie je veľa rozumnosti.

Ešte aj medzi tými niekoľkými, ktorí na mňa robia dojem náznakom svitajúcej impozantnosti – nemyslím si, že by sa ich majstrovstvo v rozumnosti dalo porovnať povedzme s majstrovstvom Johna Conwaya v matematike. Základné vedomosti, z ktorých vychádzame pri budovaní nášho chápania – keby ste vybrali iba tie časti, ktoré naozaj používame, a nie všetko, čo sme museli preštudovať, aby sme to našli – sa pravdepodobne nedajú porovnať s tým, čo vie profesionálny jadrový inžinier o jadrovom inžinierstve. Možno sa to dokonca nedá porovnať ani s tým, čo vie stavebný inžinier o mostoch. Trénujeme svoje zručnosti, to áno, tými ad-hoc spôsobmi, ktoré sme sa sami naučili; ale takýto tréning sa asi nedá porovnať v tréningovým režimom, cez aký prejde olympijský bežec, alebo hoci len obyčajný profesionálny hráč tenisu.

A koreňom *tohto* problému je, myslím si, že sme zatiaľ naozaj nedali dokopy a nesystematizovali naše zručnosti. Museli sme si toto všetko vytvoriť sami pre seba, ad hoc, a existuje hranica, koľko toho dokáže urobiť jedna myseľ, dokonca aj keď sa jej darí ťažiť z práce urobenej v iných odvetviach.

Hlavnou prekážkou k robenou tohto tak, ako by sa to *naozaj* malo robiť, je náročnosť testovania *výsledkov* programov výcviku rozumnosti, aby sme mohli mať výcvikové metódy založené na indíciách. Ešte o tom budem písať, pretože si myslím, že rozoznať úspešný výcvik a odlíšiť ho od neúspechu je podstatná, blokujúca prekážka.

Z času na čas sa robia pokusy ohľadom zásahov na odstránenie niektorých konkrétnych skreslení, ale zvykne to byť niečo ako: „Nechajte študentov hodinu cvičiť toto, potom ich otestujte o dva týždne neskôr.“ Nie: „Nechajte polovicu prihlásených prejsť verziou A trojmesačného letného výcvikového programu, polovicu verziou B, a urobte prieskum o päť rokov neskôr.“ Tu vidíte množstvo úsilia, o ktorom si myslím, že by sa vynaložilo na výcvikový program pre ľudí, ktorí by brali rozumnosť Naozaj Vážne, na rozdiel od sklonu robiť náhodné výstrely od boku, ktoré si vyžadujú asi hodinku úsilia alebo niečo také.

Daniel Burfoot má skvelú myšlienku, že toto je dôvod, prečo sa inteligencia zdá byť takým veľkým faktorom v rozumnosti – že keď všetko improvizujete ad hoc, a máte veľmi málo tréningu alebo systematickej praxe, inteligencia skončí ako tej najdôležitejší faktor v tom, čo zostalo.

Prečo „racionalistov“ neobklopuje viditeľná aura impozantnosti? Prečo sa nenachádzajú na vrchných úrovniach každej elity vybranej na hocijakom základe, ktorý má niečo spoločné s myslením? Prečo väčšina „racionalistov“ vyzerá iba ako obyčajní ľudia, možno mierne nadpriemerne inteligentní, s jedným koníčkom navyše?

Na toto existuje niekoľko odpovedí; ale jedna z nich je iste tá, že dostali menej systematického výcviku rozumnosti v menej systematickom kontexte než má prvý dan s čiernym opaskom v búchaní do ľudí.

Sám seba z tejto kritiky nevynímam. Nie som žiaden beisutsukai, pretože existujú hranice, koľko veľa Umenia dokážete vytvoriť sám, a ako dobre môžete hádať, keď nemáte podložené štatistiky výsledkov. Viem o *jednom* použití rozumnosti, ktoré by sa dalo nazvať „redukcia mätúcich poznatkov“. Toto som žiadal od svojho mozgu, toto mi dal. Myslím si, že existujú aj iné umenia, ktorých učenie by vyspelý program tréningu rozumnosti nevynechal, ktoré by ma urobili silnejším a šťastnejším a efektívnejším – keby som len mohol prejsť štandardizovaným tréningovým programom používajúcim tie najlepšie vzdelávacie techniky, ktorých účinnosť sa experimentálne preukázala. Ale ten druh ohromného, sústredeného úsilia, aké vkladám do vytvorenia môjho jedného *pod-umenia* rozumnosti od nuly – v mojom živote nie je miesto na viac než jednu takúto vec.

Považujem sa za viac než prvý dan s čiernym opaskom, a zároveň menej. Dokážem *päsťou* preraziť tehlu, a pracujem na tom, aby to postupne bola oceľ a nakoniec adamantium, ale máme iba skúsenosti občasného pouličného bitkára, ako sa kope alebo hádže alebo blokuje.

Prečo existujú školy bojových umení, ale nie dódžó rozumnosti? (Toto bola prvá otázka, ktorú som sa opýtal vo svojom prvom článku na blogu.) Vari je dôležitejšie biť ľudí než myslieť?

Nie, ale je jednoduchšie overiť si, kedy *ste* niekoho udreli. To je časť dôvodu, veľmi ústredná časť.

Ale možno ešte dôležitejšie – existujú ľudia, ktorí sa *chcú* biť, a ktorí majú predstavu, že by malo existovať systematické umenie bitky, ktoré z vás urobí viditeľne impozantnejšieho bojovníka, s rýchlosťou a eleganciou a silou za hraniami úsilia netrénovaných. Idú teda do školy, ktorá im sľubuje, že ich to naučí. A tieto školy existujú, pretože pred mnohými rokmi mali niektorí ľudia pocit, že sa dá aj viac. A tak sa dali dokopy a vymenili si svoje techniky a cvičili a formalizovali a cvičili a vyvinuli Systematické Umenie Bitky. Dotlačili sa tak ďaleko, pretože *si mysleli, že by mali byť úžasnejší* a boli ochotní na tom *makať*.

A potom... sa niekam s touto snahou *dostali*, na rozdiel od tisícov iných, ktorí sa snažili o úspech a zlyhali, pretože dokázali *rozoznať*, kedy niekoho zasiahli; a tak školy proti sebe navzájom súťažili v realistických súbojoch s jasne definovanými víťazmi.

Ale ešte predtým... najprv existovala snaha, želanie stať sa silnejším, pocit, že sa dá aj viac. Predstava rýchlosti a elegancie a sily, ktorú ešte nemali, aby *mohli* by mať, *keby* boli ochotní dať do toho veľa práce, ktorá ich viedla k systematizovaniu a tréningu a testovaniu.

Prečo nemáme Umenie Rozumnosti?

Po tretie, pretože dnešní „racionalisti“ majú problém pracovať v skupinách: o tomto ešte poviem viac.

Po druhé, pretože je ťažké overiť úspech v tréningu, alebo ktorá z dvoch škôl je silnejšia.

Ale po prvé, pretože ľuďom chýba ten pocit, že rozumnosť je niečo, čo by sa *malo* systematizovať a trénovať a testovať ako bojové umenie, a čo by malo mať za sebou toľko vedomostí ako jadrové inžinierstvo, čo by superhviezdy mali cvičiť tak tvrdo ako šachoví veľmajstri, čoho úspešní cvičenci by boli obklopení zrejmou aurou úžasnosti.

A takisto sa neobzreli na *nedostatok* viditeľne väčšej impozantnosti a nepovedali si: „Určite niečo robíme zle.“

„Rozumnosť“ skrátka vyzerá ako nejaký koníček navyše, o ktorom ľudia hovoria na večierkoch; prijatý režim konverzačnej pózy s málo ak vôbec nejakými skutočnými dôsledkami; a ani sa nezdá, že by na tomto bolo niečo zle.

*→

314. Epistemická skazenosť

Niekto si za toto zaslúži veľké poďakovanie, ale už si nespomínam, kto; moje záznamy neobsahujú žiaden e-mail ani komentár na OB, ktorý mi povedal o tejto 12-stránkovej eseji „Epistemická skazenosť v bojových umeniach“ od Gilliana Russella.[297] Žeby Anna Salamon?

> Všetci sme sa zoradili vo svojich kravatách a elegantných topánkach (toto bolo Anglicko) a napodobňovali sme ho – ľavá, pravá, ľavá, pravá – a potom nám povedal, že keby sme cvičili vo vzduchu dostatočne oddane celé tri roky, potom by sme dokázali svojimi údermi zabiť býka jedinou ranou.
>
> Uctievala som pána Howarda (hoci by som radšej zomrela než mu to povedala) a tak som chudé jedenásťročné dievčatko začala veriť, že keby som cvičila, dokázala by som zabiť býka jedinou ranou, ked budem mať štrnásť.
>
> Táto esej je o epistemickej skazenosti v bojových umeniach, a tento príbeh ilustruje presne to. Hoci slovo „skazenosť“ zvyčajne naznačuje úmyselnú krutosť a násilie, ja ho tu budem používať v staromódnejšom význame, niečo plné kazov.

Všetko je tam *nádherne* zovšeobecnené. Aby som zhrnul niektoré z kľúčových pozorovaní, ako vzniká epistemická skazenosť:

→ http://lesswrong.com/lw/2c/a_sense_that_more_is_possible/

297 Gillian Russell, „Epistemic Viciousness in the Martial Arts,“ in *Martial Arts and Philosophy: Beating and Nothingness*, ed. Graham Priest and Damon A. Young (Open Court, 2010).

- Umenie, dódžó a sensei sú vnímaní ako posvätní. „Mať červené nechty v dódžó je ako ísť do kostola v minisukni a s hlbokým výstrihom... O študentoch iných bojových umení sa hovorí, akoby praktizovali nesprávne náboženstvo.“
- Ak vás váš učiteľ odvedie nabok a naučí vás špeciálny pohyb a vy ho cvičíte dvadsať rokov, veľa ste do toho emocionálne investovali, a budete chcieť ignorovať všetky prichádzajúce indície proti tomuto pohybu.
- Prichádzajúci študenti nemajú príliš na výber: bojové umenie sa nedá naučiť z knihy, takže musia dôverovať učiteľovi.
- Úcta voči známym historickým majstrom. „Bežci si myslia, že dnešná reakcia Sveta bežcov vie o behaní viac než všetci starovekí Gréci dokopy. A neplatí to iba pre beh, alebo iné fyzické aktivity, že história má vyhradené svoje miesto; to isté platí aj v ľubovoľnej rozvinutej oblasti štúdia. Nepovažuje sa za neúctivé, ak nejaký fyzik povie, že teórie Isaaca Newtona sú nepravdivé...“ (Znie vám to povedome?)
- „My, bojoví umelci, zápasíme s istou formou chudoby – chudoby údajov – ktorá spôsobuje, že naše názory je ťažko testovať... Pokiaľ nemáte dostatočnú smolu na to, aby ste bojovali vojnu holými rukami, nemôžete si *overiť*, koľko presne sily a v akom presne uhle treba na zlomenie krku...“
- „Ak nemôžete otestovať efektivitu nejakej techniky, potom je ťažké testovať metódy na zlepšenie tejto techniky. Mali by ste cvičiť svoje nukite vo vzduchu, alebo vás to iba povzbudí príliš vystierať ruku? … Naša neschopnosť otestovať si svoje bojové metódy obmedzuje našu schopnosť otestovať si svoje výcvikové metódy.“
- „Ale skutočným problémom nie je len to, že žijeme v chudobe údajov – myslím si, že to platí aj pre niektoré úctyhodné disciplíny, vrátane teoretickej fyziky – problém je, že žijeme v chudobe, ale naďalej sa správame, akoby sme žili v luxuse, akoby sme si mohli bezpečne dovoliť veriť hocičomu, čo nám povedia...“ **(+10!)**

Jedna vec, ktorú som si pamätal z tejto eseje, ale po druhom prečítaní sa ukázalo, že to tam v skutočnosti nie je, bolo degenerovanie bojových umení po poklese skutočných bojov – čím myslím boje, kde sa ľudia naozaj snažili jeden druhého zraniť a občas niekto zomrel.

V tých dňoch ste mali nejakú predstavu, kto sú skutoční majstri a ktorá škola dokáže poraziť ktorú.

A potom sa veci *scivilizovali*. A potom šlo všetko dole kopcom až do bodu, kde máme na YouTube videá, kde údajný čierny opasok s N-tým danom dostane výprask od niekoho, kto má skúsenosť s naozajstnou bitkou.

Počul som o jednom takomto prípade, ktorý bol *naozaj* smutný; bol to majster nejakej školy, ktorý bol presvedčený, že dokáže použiť techniky *čchi*. Jeho študenti naozaj spadli, keď na nich použil útok čchi, čo je zláštny a pozoruhodný prípade sebahypnózy alebo *niečoho*... a potom sa majster postavil proti skeptikovi a samozrejme dostal totálny výprask.

Pravdivo sa hovorí, že „vedieť, ako neprehrať“ je širšie uplatniteľná informácia než „ako vyhrať“. Každý jeden z týchto rizikových faktorov sa priamo prenáša do ľubovoľného pokusu začať „dódžó rozumnosti“. Kladiem vám otázku: Čo sa s tým dá robiť?

* →

315. Bez indícií sa školy množia

Robyn Dawes, autor jedného z pôvodných článkov zo zbierky *Usudzovanie pri neistote* a knihy *Rozumná voľba v neistom svete* – jeden z mála, ktorí s naozaj tvrdo snažili importovať výsledky do skutočného života – je zároveň autorom knihy *Dom z kariet: Psychológia a psychoterapia postavené na piesku.*

→ http://lesswrong.com/lw/2i/epistemic_viciousness/

Z *Domu z kariet*, kapitoly 1:[298]

> Schopnosti týchto profesionálov boli podrobené empirickému preskúmaniu – napríklad ich efektivita ako terapeutov (Kapitola 2), ich vhľady o ľuďoch (Kapitola 3), a vzťah medzi tým, ako dobre fungujú a koľko indície majú vo svojom odvetví (Kapitola 4). Prakticky všetky výskumy – a táto kniha odkazuje na vyše tristo empirických výskumov a sumárov výskumov – zistili, že tvrdenia týchto profesionálov o ich lepšom intuitívnom vhľade, porozumení, a terapeutickej schopnosti sú jednoducho neplatné.

Pamätáte sa na Rorschachov test s atramentovými škvrnami? Je to veľmi pôsobivý argument: pacient sa pozrie na atramentovú štvrnu a povie, čo vidí, psychoterapeut na základe toho interpretuje jeho psychologický stav. Stovky experimentov hľadali nejakú indíciu, že to naozaj funguje. Keďže čítate tieto slová, viete už uhádnuť, že odpoveď je jednoducho: „Nie." Napriek tomu sa Rorschach stále používa. Je to jednoducho taký *dobrý príbeh*, že sa psychoterapeuti nedokážu primäť uveriť obrovským horám experimentálnej indície, ktorá hovorí, že to nefunguje...

...čo vám povie, a akým druhom odvetvia tu máme do činenia.

A experimentálne výsledky v tejto oblasti ako celku sú porovnateľné. Áno, vie sa, že stav pacientov, ktorí sa stretávajú s psychoterapeutmi, sa zlepšuje rýchlejšie než pacientov, ktorí jednoducho nerobia nič. Neexistuje však štatisticky merateľný rozdiel medzi mnohými školami psychoterapie. Neexistuje žiaden merateľný rozdiel, ktorý získate rokmi odbornosti.

A neexistuje ani merateľný rozdiel medzi stretnutím sa s psychoterapeutom a strávením rovnakého množstva času rozhovorom s náhodne vybranými vysokoškolským učiteľom z inej oblasti. Zdá sa, že vám skrátka pomáha, keď sa rozprávate s *hocikým*.

V tejto úplnej neprítomnosti štipky experimentálnej indície ich efektivity, psychoterapeuti dostávajú od štátu licencie, ich svedctvo sa prijíma na súde, ich školy dostávajú akreditácie, a ich šeky platí zdravotné poistenie.

A v psychoterapii takisto bolo obrovské rozmnoženie „škôl", alebo tradícií praxe; napriek – alebo azda *kvôli* – nedostatku ľubovoľných experimentov, ktoré by ukázali, že jedna škola je lepšia než druhá...

Naozaj by som raz mal napísať viac o všetkých tých smutných veciach, ktoré toto vypovedá o našom svete; o tom, ako *podstata medicíny*, ako ju rozoznáva spoločnosť a súdy, nie je zbierkou procedúr so štatistickými indíciami pre ich účinnosť liečenia; ale skôr, tá správna atmosféra autority.

Ale dnešná téma je množenie sa tradícií v psychoterapii. Pokiaľ viem, toto je spôsob, ako vo svojej oblasti môžete nadobudnúť prestíž – nie tým, že objavíte úžasnú novú techniku, ktorej účinnosť by sa dala experimentálne overiť a používať všetkými; ale skôr tým, že oddelíte svoju vlastnú „školu", podloženú vašou charizmou zakladateľa, a dobrými príbehmi, ktoré ste porozprávali o všetkých dôvodoch, prečo by vaše techniky *mali* fungovať.

Toto pravdepodobne v nemalej miere zodpovedá za vôbec existenciu a pokračovanie psychoterapie – prísľub, že sa z vás stane Majster, ako bol Freud, ktorý to urobil ako prvý (tiež bez najmenšej štipky experimentálnej indície). Toto je ten mosadzný prsteň úspechu, ktorý možno naháňať – vyhliadka na to, že budete guru a budete mať svojich vlastných stúpencov. Je to zápas o stúpencov, čo udržiava duchovenstvo pri živote.

Toto sa stane, keď sa nejaká oblasť odpúta od experimentálnych indícií – hoci existujú aj iné faktory, ktoré robia psychoterapiu rizikovou, napríklad úcta, ktorú im preukazujú ich pacienti, túžba spoločnosti veriť, že mentálne uzdravovanie je možné, a samozrejme všeobecné nebezpečenstvo súvisiace s hovorením ľuďom, ako majú myslieť.

(Dawes toto napísal v 80-tych rokoch, a ja viem, že Rorschach sa používal ešte v 90-tych, ale je možné, že sa odvtedy veci zlepšili (ako tvrdí jeden diskutér). Spomínam si, že som počul, že existuje pozitívna indícia pre väčšiu účinnosť kognitívne behaviorálnej terapie.)

Oblasť hedonickej psychológie (štúdium šťastia) vznikla do istej miery na základe uvedomenia si, že šťastie sa dá *merať* – existovala skupina mier, ktoré sa, ježkove oči, navzájom dobre validovali.

298 Robyn M. Dawes, *House of Cards: Psychology and Psychotherapy Built on Myth* (Free Press, 1996).

Akt vytvorenia nového merania vytvára novú vedu; ak je to *dobré* meranie, dostanete dobrú vedu.

Ak chcete vytvoriť a zorganizovať prax hocičoho, naozaj potrebujete nejaký spôsob, ako povedať, ako dobre to robíte, a praktizovať seriózne testovanie – to chce kontrolnú skupinu, experimentálnu skupinu, a štatistiku – dôveryhodne vyzerajúcich techník, ktoré ľudia vymyslia. *Naozaj* to potrebujete.

⁂→

316. 3 úrovne overovania rozumnosti

Mám silné podozrenie, že môže existovať umenie rozumnosti (dosahovanie mapy, ktorá odráža územie, rozhodovanie, ktoré smeruje skutočnosť do oblastí vysoko vo vašom rebríčku hodnôt), ktoré ide ďalej než štandardné zručnosti, a ďalej než to, čo vie ľubovoľný jeden praktizujúci. Mám pocit, že sa dá viac.

Do akej miery s tým nejaká *skupina* ľudí dokáže urobiť niečo užitočné, bude závisieť *v prevažnej väčšine* od toho, aké metódy dokážeme vymyslieť na *overovanie* našich mnohých úžasne dobrých nápadov.

Navrhujem rozdeliť metódy overovania na 3 úrovne užitočnosti:

- Založené na povesti
- Experimentálne
- Organizačné

Ak váš majster bojového umenia občas bojuje v realistických dueloch (ideálne v *skutočných* dueloch) proti majstrom iných škôl, a vyhráva alebo prinajmenšom neprehráva príliš často, potom viete, že majstrova *povesť* je *ukotvená v skutočnosti*; viete, že váš majster nie je úplný pozér. To isté platí, ak vaša škola pravidelne súťaží proti iným školám. *Držíte sa skutočnosti.*

Niektoré bojové umenia nesúťažia dostatočne realisticky, a ich študenti idú k zemi za pár sekúnd proti skutočným pouličným bitkárom. Iné školy bojových umení nesúťažia *vôbec* – nanajvýš charizmou a dobrými príbehmi – a ich majstri zisťujú, že majú schopnosti čchi. Do tejto druhej skupiny by sme mohli zaradiť aj roztrieštené školy psychoanalýzy.

Takže dokonca aj ten základný krok, pokúsiť sa *ukotviť povesť* v nejakom realistickom teste inom než je charizma a dobré príbehy, má ohromný pozitívny dopad na celú vašu oblasť snahy.

Ale to vám ešte nedáva vedu. Veda vyžaduje, aby ste dokázali testovať 100 použití metódy A proti 100 použitiam metódy B a zbehli štatistiku na výsledkoch. *Experimenty* majú byť replikovateľné a replikované. To si vyžaduje *štandardné merania*, ktoré možno zbehnúť na študentoch, ktorí sa učili náhodne pridelené alternatívne metódy, nie iba *realistické duely* vybojované medzi majstrami používajúcimi všetky svoje nahromadené techniky a silu.

Oblasť štúdia šťastia vznikla viacmenej tak, že sme si uvedomili, že opýtať sa ľudí: „Na škále od 1 do 10, ako dobre sa práve teraz cítite?“ je miera, ktorá sa štatisticky dobre validuje voči iným predstavám merania šťastia. A toto, napriek všetkému skepticizmu, vyzerá ako pomerne užitočná miera nejakých vecí, ak sa opýtate 100 ľudí a spriemerujete výsledky.

Ale predstavte si, že by ste chceli, aby sa najšťastnejší ľudia dostali do mocenských pozícií – chcete platiť šťastných ľudí, aby cvičili druhých ľudí, ako byť šťastnejšími, alebo chcete zamestnať tých najšťastnejších v hedžovom fonde? Potom potrebujete nejaký test, ktorý je *ťažšie obabrať* než keď sa niekoho iba opýtate: „Aký ste šťastný?“

Otázka overovacích metód, ktoré sú dosť dobre na to, aby sa na nich vybudovali *organizácie*, je obrovský problém na všetkých úrovniach modernej spoločnosti. Ak chcete používať SAT, aby ste riadili prijímačky na elitné vysoké školy, dá sa SAT zdolať tým, že sa budete učiť *iba* na SAT spôsobom, ktorý nakoniec nebude korelovať s iným študijným potenciálom? Ak dáte vysokým školám moc udeľovať diplomy, majú potom motív ľuďom pomáhať? (Považujem za absolútne samozrejmé, že úloha overovať získané zručnosti a teda moc udeľovať tituly by mala byť oddelená od inštitúcií, ktoré učia, ale teraz tam

→ http://lesswrong.com/lw/2j/schools_proliferating_without_evidence/

nezachádzajme.) Ak hedžový fond uverejňuje výnosnosť 20 %, naozaj sú o toľko lepší než indexy, alebo iba predávajú puty, ktoré vybuchnú, keď trh klesne?

Ak máte metódu overovania, ktorú možno obabrať, celá oblasť sa prispôsobí na jej obabrávanie, a tak stratí svoj účel. Vysoké školy sa menia na testy toho, či dokážete vydržať presedieť hodiny. Stredné školy nerobia nič iné než učia na štátne maturity. Hedžové fondy predávajú puty, aby zvýšili svoje výnosy.

Na druhej strane – stále sa nám darí učiť inžinierov, dokonca aj keď naše organizačné metódy overovania nie sú dokonalé. Takže ktoré dokonalé či nedokonalé metódy by ste mohli použiť pri overovaní zručností rozumnosti, ktoré by boli aspoň *trochu* odolné voči obabrávaniu?

(Merania s vysokým šumom možno stále *experimentálne* použiť, ak náhodne rozdelíte dosť pokusných osôb na to, aby ste v priemere neutralizovali rozptyl. Ale na *organizačné* účely overovania konkrétnych jednotlivcov potrebujete merania s nízkym šumom.)

Teraz vám teda kladiem otázku – ako overíte zručnosti rozumnosti? Na ľubovoľnej z týchto troch úrovní. Brainstormujte, prosím vás; dokonca aj náročné a drahé meranie sa môže stať zlatým štandardom na overovanie iných meraní. Nech sa páči, pošlite mi e-mail na yudkowsky@gmail.com a navrhnite ľubovoľné merania, pre ktoré je lepšie, ak sa o nich verejne nevie (hoci je to samozrejme veľká nevýhoda danej metódy). Aj hlúpe nápady môžu viesť k dobrým nápadom, takže ak neviete vymyslieť dobrý nápad, *vymyslite nejaký hlúpy*.

Založené na povesti, experimentálne, organizačné:

- Niečo, čo môžu robiť majstri a školy, aby sa držali skutočnosti (realistickej skutočnosti);
- Niečo, čo môžete odmerať u každého zo sto študentov;
- Niečo, čo môžete použiť ako test ešte aj vtedy, keď majú ľudia motív to obabrať.

Nájdenie dobrých riešení na každej úrovni rozhoduje o tom, na čo môže byť celé odvetvie štúdia užitočné – ako veľa môže dúfať, že dosiahne. Toto je jedna z Veľkých Dôležitých Základných Otázok, takže...

Myslite!

(**PS:** A uvažujte samostatne predtým než sa pozriete na nápady druhých; potrebujeme široké pokrytie.)

*→

317. Prečo náš druh nedokáže spolupracovať

Z čias, keď som tam ešte musel chodiť, si pamätám, ako v našej synagóge vyzývali na každoročné finančné dary. Ak si dobre spomínam, malo to pomerne jednoduchý formát. Rabín a pokladník hovorili o výdavkoch synagógy a aká dôležitá je táto každoročná zbierka, a potom členovia synagógy zo svojich miest vykrikovali svoje záväzky.

Priamočiare, nie?

Teraz vám porozprávam o inej výzve na každoročné finančné dary. O jednej, ktorú som naozaj organizoval, počas raných rokoch Výskumného ústavu strojovej inteligencie. Jeden rozdiel bol v tom, že sa táto výzva konala cez internet. A druhý rozdiel bol v tom, že obecenstvo prevažne pochádzalo z davu ateistov / libertariánov / technofilov / fanúšikov sci-fi / inovátorov / programátorov / atď. (Aby som ukázal približným smerom na empirický zhluk v priestore osôb. Ak rozumiete slovnému spojeniu „empirický zhluk v priestore osôb“, potom viete, o kom hovorím.)

Výzvu na finančné dary som zostavoval veľmi opatrne. Zo svojej povahy som príliš hrdý na to, aby som žiadal druhých o pomoc; ale vyše 60 % tohto zdráhania som rokmi prekonal. Táto neziskovka potrebovala peniaze a rástla príliš pomaly, tak som do tohoročnej výzvy vložil trochu sily a poézie. Poslal som ju do niekoľkých mailových skupín, ktoré pokrývali väčšinu našej potenciálnej sponzorskej základne.

→ http://lesswrong.com/lw/2s/3_levels_of_rationality_verification/

A takmer okamžite začali ľudia písať do týchto mailových skupín o dôvodoch, prečo nedarujú. Niektorí z nich nadniesli základné otázky o filozofii a poslaní tejto neziskovky. Iní hovoril o svojich skvelých nápadoch, z akých *iných* zdrojov by táto neziskovka mohla dostať peniaze, namiesto od nich. (Neponúkali sa ako dobrovoľníci, že *sami* oslovia niektoré z týchto zdrojov, akurát mali nápady, ako by sme to mohli urobiť *my*.)

Teraz asi poviete: „No, možno vaše poslanie a filozofia *mali* základné problémy – nechcel by si predsa *cenzurovať* takúto diskusiu, alebo áno?“

Zapamätajte si túto myšlienku.

Pretože ľudia *darovali*. Začali sme dostávať dary ihneď, cez Paypal. Dokonca sme dostali blahoprajné poznámly hovoriace, ako ich táto výzva konečne rozhýbala. Pri dare 111,11 dolárov bola priložená správa: „Rozhodol som sa dať o čosi viac. O jednu stovku, o jednu desiatku, o jeden dolár, o jeden desaťcent, a o jeden cent. Možno nie sú všetci za jedného, ale tento jeden sa snaží byť za všetkých.“

Lenže žiaden z týchto darcov neuverejnil svoj súhlas v mailovej skupine. Ani jeden.

Takže pokiaľ mohli títo darcovia vedieť, boli sami. A keď sa vrátili nasledujúci deň, neobjavili poďakovanie, ale argumenty, prečo *nemali* darovať. Kritika, zdôvodnenie nedarovania – *iba tieto* boli hrdo verejne vystavené.

Akoby pokladník dokončil svoju každoročnú výzvu, a každý, kto si *nedal* záväzok, by hrdo vstal a vykríkol svoj dôvod, prečo odmieta; zatiaľ čo tí, ktorí si dali záväzok, by ho zašepkali potichu, aby nikto nepočul.

Poznám niekoho, kto má racionalistické poslanie, a chodí sa žalostne pýtať: „Ako je možné, že ufologická sekta Raeliánov dokáže získať desaťtisíce členov [pravdepodobne okolo 40 000] pre úplný nezmysel, ale my nedokážeme zohnať ani tisíc ľudí, aby nám pomohli s týmto?“

Samozrejmý nesprávny spôsob, ako túto myšlienku dokončiť, je: „Urobme to isté čo Raeliáni! Pridajme do tohto mému trochu nezmyslov!“ Pre dobro tých, ktorých okamžite nezastavili ich etické zábrany, poznamenám, že na každú jednu slávnu ufologickú sektu pripadá sto takých, ktoré zlyhali. Aj Temná Strana si môže vyžadovať zriedkavé schopnosti, ktoré *vy*, áno *vy*, nemáte: Nie každý má na to, aby bol vládca Sithov. Konkrétne, ak o svojich plánovaných lžiach diskutujete verejne na internete, tak zlyháte. Nie som žiaden majster zločinu, ale aj ja dokážem posúdiť, že niektorí ľudia nemajú na to, aby boli podvodníkmi.

Takže asi nie je dobrá myšlienka pestovať si pocit porušených nárokov pri myšlienke, že nejaká *iná* skupina, o ktorej si myslíte, že je oproti vám *podradná*, má viac peňazí a nasledovníkov. Táto cesta vedie – pardon za výraz – k Temnej Strane.

Ale pravdepodobne *má* zmysel, aby sme začali klásť sami sebe pár ostrých otázok, ak sa údajní „racionalisti“ nedokážu *skoordinovať* ani tak dobre ako ufologická sekta.

Ako to funguje na Temnej Strane?

Uznávaný vodca hovorí, a nasleduje zbor číreho súhlasu: ak je tam niekto, kto v sebe prechováva pochybnosti, nechá si ich pre seba. Všetci jednotlivci v obecenstve teda vidia túto atmosféru číreho súhlasu, a cítia väčšiu dôveru v predložené myšlienky – dokonca aj keď oni osobne v sebe prechovávajú pochybnosti, ale všetci *ostatní* vyzerajú, že s tým súhlasia.

(Štandardné označenie na toto je „pluralistická nevedomosť.“)

Ak je niekto aj po tomto stále nepresvedčený, opustí skupinu (alebo na niektorých miestach ho popravia) – a zvyšok súhlasí viac, a posilňuje sa navzájom s menším rušením.

(Toto som nazval „ochladzovanie skupín vyparovaním“.)

Samotné *myšlienky*, nie iba vodca, vytvárajú neobmedzené nadšenie a chválu. Efekt svätožiary je, že vnímanie všetkých pozitívnych vlastností koreluje – napríklad ak poviete pokusným osobám, že nejaký konzervačný prostriedok má nejaké výhody, budú ho hodnotiť ako menej riskantný, aj keď tieto údaje nazvájom logicky nesúvisia. Toto dokáže vytvoriť efekt pozitívnej spätnej väzby, keď nejaká myšlienka vyzerá lepšie a lepšie a lepšie, najmä ak je kritika vnímaná ako zrada a hriech.

(Čo nazývam „afektívna špirála smrti“.)

Toto sú teda všetko príklady silných síl Temnej Strany, ktoré dokážu stmeliť skupiny.

A predpokladám, že *my* by sme nezašli tak ďaleko, aby sme zašpinili ruky týmto...

Preto, ako skupina, bude Svetlá Strana vždy rozdelená a slabá. Ateisti, libertariáni, technofili, kockáči, fanúšikovia sci-fi, vedci, ba dokonca aj nefundamentalistické náboženstvá, nebudú nikdy schopní konať s fanatickou jednotou, ktorá oživuje radikálny islam. Technologické výhody siahajú iba odtiaľ potiaľ; vaše nástroje možno skopírovať alebo ukradnúť a použiť proti vám. V konečnom dôsledku Svetlá Strana vždy prehrá v každom skupinovom konflikte, a budúcnosť nevyhnutne patrí Temným.

Myslím si, že reakcia človeka na túto vyhliadku povie veľa o jeho postoji k „rozumnosti".

Niektorí pisatelia o „Zrážke Civilizácií" vyzerajú zmierení s tým, že osvietenie má predurčené z dlhodobého hľadiska prehrať voči radikálnemu islamu, a vzdychajú, a smutne kývajú hlavami. Predpokladám, že tým chcú signalizovať svoju cynickú sofistikovanosť alebo také niečo.

Ja som si osobne vždy myslel – povedzte, že mi šibe – že *skutočný* racionalista by mal byť *efektívny v skutočnom svete*.

Takže ja mám problém s predstavou, že Temná Strana vďaka svojej *pluralistickej nevedomosti a afektívnym špirálam smrti* vždy vyhrá, pretože sa *skoordinujú lepšie* než my.

Mohli by ste si azda pomyslieť, že *skutoční* racionalisti by mali byť *viac* koordinovaní? Iste všetok ten nerozum musí mať aj nejaké *nevýhody*? Ten režim nemôže byť *optimálny*, alebo áno?

A ak sa súčasné „racionalistické" skupiny *nedokážu* skoordinovať – ak nedokážu podporiť skupinové projekty rovnako dobre ako jediná synagóga vyberá dary od svojich členov – no, nechám na vás, aby ste dokončili tento sylogizmus.

Existuje porekadlo, ktoré občas používam: „Je nebezpečné byť polovičným racionalistom."

Napríklad si viem predstaviť spôsob, ako podkopať niekoho inteligenciu tým, že ho *selektívne* naučím niektoré metódy rozumnosti. Predstavte si, že niekoho naučíte dlhý zoznam logických omylov a kognitívnych skreslení, a vycvičíte ho všímať si tieto omyly a skreslenia v argumentoch druhých ľudí. Ale vybrali ste si opatrne tie omyly a skreslenia, z ktorých je *najľahšie obviňovať* druhých, tie najvšeobecnejšie, ktoré sa dajú ľahko nesprávne použiť. A *neupozorníte* ich, aby prešetrovali *argumenty, s ktorými súhlasia*, rovnako tvrdo ako prešetrujú chyby v *nesúhlasiacich* argumentoch. Získali teda veľký reprertoár chýb, z ktorých budú obviňovať iba argumenty a argumentátorov, ktorí sa im nepáčia. Mám podozrenie, že toto je jedna z hlavných ciest, ako sa bystrí ľudia stávajú hlúpymi. (A všimnite si, mimochodom, že som som vám práve dal ďalší Plne Všeobecný Protiargument proti bystrým ľuďom, ktorých argumenty sa vám nepáčia.)

Podobne, keby ste chceli zaručiť, že skupina „racionalistov" nikdy nezvládne žiadnu úlohu, ktorá si vyžaduje viac než jedného človeka, mohli by ste ich naučiť iba techniky individuálnej rozumnosti, bez spomenutia čohokoľvek o technikách koordinovanej skupinovej rozumnosti.

Neskôr napíšem viac o tom, ako si myslím, že by sa racionalisti mohli vedieť lepšie koordinovať. Ale dnes sa chcem sústrediť na to, čo by ste mohli nazvať *kultúrou nesúhlasu*, alebo dokonca *kultúrou námietok*, čo jedna z dvoch hlavných síl, ktoré bránia davu technofilov skoordinovať sa.

Predstavte si, že ste na konferencii, a rečník urobí 30-minútovú prednášku. Potom sa ľudia zoradia pri mikrofónoch s otázkami. Prvá otázka je námietka voči grafu s logaritmickou škálou použitému na obrázku 14; cituje *Vizuálne zobrazenie kvantitatívnej informácie* od Tufteho. Druhá otázka spochybňuje tvrdenie na obrázku 3. Tretia otázka je návrh alternatívnej hypotézy, ktorá napohľad vsvetľuje tie isté údaje...

Dokonale normálne, však? Teraz si predstavte, že ste na konferencii, a rečník urobí 30-minútovú prednášku. Ľudia sa zoradia pri mikrofóne.

Prvý človek povie: „Súhlasím si všetkým, čo si v prednáške povedal, a myslím si, že si skvelý." Potom odstúpi.

Druhý človek povie: „Obrázok číslo 14 bol krásny, veľa som sa z neho naučil. Si úžasný." Odstúpi.

Tretí človek...

No, nikdy sa nedozviete, čo chcel do mikrofónu povedať ten tretí, pretože v tom čase už s výkrikom utekáte preč z miestnosti, poháňaný hrôzou do špiku kostí, akoby na pódium vyskočil Cthulhu, strachom z nemožno neprirodzeného javu, ktorý vtrhol na vašu konferenciu.

Áno, skupina, ktorá nedokáže tolerovať nesúhlas, nie je rozumná. Ale ak tolerujete *iba* nesúhlas – ak tolerujete nesúhlas, *ale nie súhlas* – potom tiež nie ste rozumní. Ste ochotní počuť iba niektoré úprimné názory, ale nie iné. Ste nebezpečný polovičatý racionalista.

Cítime sa *blízko* seba rovnako nepohodlne ako sa členovia ufologickej sekty cítia *ďaleko* od seba. Ani to nemôže byť správne. Opak hlúposti nie je inteligencia.

Povedzme, že máme dve skupiny vojakov. V skupine 1, pešiaci nevedia o taktike a stratégii; iba seržanti vedia o taktike, a iba dôstojníci vedia o stratégii. V skupine 2 každý na každej úrovni vie všetko o taktike a stratégii.

Mali by sme očakávať, že skupina 1 porazí skupinu 2, pretože skupina 1 sa bude riadiť príkazmi, zatiaľ čo každý v skupine 2 príde s *lepším nápadom* než bol hocijaký rozkaz, ktorý dostal?

V tomto prípade by som musel pochybovať o tom, nakoľko skupina 2 naozaj rozumie vojenskej teórii, pretože je *základnou* poučkou, že nekoordinovaný dav bude zmasakrovaný.

Ak sa vám darí tým horšie, čím *viac vedomostí* máte, potom robíte niečo veľmi nesprávne. Mali by ste byť vždy schopní uplatniť *prinajmenšom* tú istú stratégiu ako keby ste nič nevedeli, a pokiaľ možno *lepšiu*. Rozhodne by sa vám nemalo dariť *horšie*. Ak sa pristihnete, že ľutujete svoju „rozumnosť“, potom by ste mali prehodnotiť, čo je rozumné.

Na druhej strane, ak ste iba polovičatý racionalista, *ľahko* sa vám s väčším množstvom vedomostí môže dariť horšie. Spomínam si na krásny pokus, ktorý ukázal, že študenti so silnými politickými názormi, ktorí mali viac vedomostí o danej téme, reagovali menej na nesúhlasné indície, pretože mali viac munície, ktorou mohli protiargumentovať iba voči nesúhlasiacim indíciám.

Zdá sa, že sme sa zasekli v hroznom údolí čiastočnej rozumnosti, kde sme nakonec horšie koordinovaní než náboženskí fundamentalisti, a schopní vynaložiť menej úsilia než ufologické sekty. Iste, to malé úsilie, ktoré sa nám *podarí* vynaložiť, môže byť lepšie zamerané na pomáhanie ľuďom namiesto opaku – ale to nie je prijateľná výhovorka.

Keby sme sa podujali na systematický výcvik racionalistov, boli by tam lekcie o tom, ako nesúhlasiť, a lekcie o tom, ako súhlasiť, lekcie zamerané na to, aby sa cvičenci cítili pohodlne pri námietke, a lekcie zamerané na to, aby sa cítili pohodlne v zhode. Jeden deň príde každý oblečený inak, iný deň prídu všetci v uniforme. Musíte mať pokryté obe strany, inak ste iba polovičatý racionalista.

Dokážete si predstaviť tréning perspektívnych racionalistov, kde by nosili uniformu a synchronizovane pochodovali, a tréningové sedenia, kde by jeden s druhým súhlasili a tlieskali všetkému, čo povie rečník na pódiu? Znie to ako nevýslovná hrôza, však, akoby niečo takéto priamo priznávalo, že sme zlá sekta. Ale prečo *nie je* okej precvičovať toto, zatiaľ čo *je* okej precvičovať *nesúhlas* s každým v dave? *Nikdy* sa nedostanete do situácie, keď by ste súhlasili s väčšinou?

Naša kultúra kladie všetok dôraz na hrdinský nesúhlas a hrdinský vzdor, a žiaden na hrdinský súhlas alebo hrdinskú skupinovú zhodu. Signalizujeme svoju vyššiu inteligenciu a svoje *členstvo v spoločenstve nekonformistov* tým, že vymýšľame chytré námietky proti argumentom druhých. Možno *preto* dav technofilov / Silicon Valley vždy zostáva v menšine, prehráva boje s menej nekonformistickými frakciami širšej spoločnosti. Nie, neprehrávame preto, lebo sme lepší, prehrávame preto, lebo naše výlučne individualistické tradície podkopávajú našu schopnosť spolupracovať.

Ďalšia veľká zložka, o ktorej si myslím, že podkopáva skupinové snahy v spoločenstve technofilov, je, že sa *hanbíme za silné pocity*. Stále máme v hlave zaseknutého Spocka ako archetyp rozumnosti, rozumnosť ako nezaujatosť. Alebo možno súvisiaca chyba, rozumnosť ako cynizmus – pokúšame sa signalizovať svoju vyššiu rozhľadenú sofistikovanosť tým, že ukazujeme, že nám záleží na veciach menej než druhým. Dávame si pozor na to, aby sme okázalo verejne pozerali zhora na tých, ktorí sú natoľko naivní, že dávajú najavo, že im na niečom silne záleží.

Nemali by ste nepríjemný pocit, keby rečník na pódiu povedal, že mu na natoľko záleží povedzme na boji proti starnutiu, že by za túto kauzu ochotne zomrel?

Ale nikde v teórii pravdepodobnosti ani v teórii rozhodovania sa nepíše, že racionalistovi nesmie na ničom záležať. Pozrel som si tie rovnice a naozaj to tam nie je.

Najlepšia neformálna definícia rozumnosti, akú som kedy počul, je „To, čo môže byť zničené pravdou, by malo byť zničené.“ Mali by sme sa snažiť cítiť emócie, ktoré zodpovedajú faktom, a nie snažiť sa necítiť žiadne emócie. Ak nejaká emócia môže byť zničená pravdou, mali by sme sa jej vzdať. Ale ak je nejaká kauza hodna úsilia, potom by sme jednoznačne mali naplno preciťovať jej dôležitosť.

Za niektoré veci sa *oplatí* zomrieť. Áno, naozaj! A ak sa nedokážeme cítiť pohodlne, keď to pripustíme, a keď počujeme, ako to hovoria druhí, potom budeme mať problém s tým, aby nám dosť *záležalo* – aj aby sme sa dosť *skoordinovali* – na vynaložení úsilia pre skupinové projekty. Musíte učiť obe strany: „To, čo môže byť zničené pravdou, by malo byť zničené“ a „To, čo pravda podporuje, by malo prekvitať.“

Počul som argument, že tabu proti emocionálnemu jazyku napríklad v odborných vedeckých článkoch, je dôležité, aby sme nechali fakty bojovať proti sebe bez rozptyľovania. To ale neznamená, že by toto tabu malo platiť všade. Myslím si, že existujú časti života, kde by sme sa mali naučiť *chváliť* silný emocionálny jazyk, výrečnosť, a poéziu. Keď je niečo, čo treba urobiť, poetická výzva to pomôže urobiť, a preto si samotná zaslúži chválu.

Potrebujeme vytrvať v našom úsilí odhaľovať *škodlivé* kauzy a *neoprávnené* odvolávky, aby neprevalcovali úlohy, ktoré naozaj treba urobiť. Potrebujete obe strany – ochotu odvrátiť sa od škodlivých káuz, aj ochotu chváliť tie užitočné; silu nenechať sa pohnúť nepodloženými tvrdeniami, aj silu nechať sa pohnúť tými podloženými.

Myslím si, že tá synagóga pri ich každoročnej výzve to robila správne, naozaj. Nešli riadok za riadkom, nestavali jednotlivcov pod reflektor, nehľadeli na nich a nepýtali sa: „A koľko darujete *vy*, pán Schwartz?“ Ľudia jednoducho ohlásili svoje záväzky – nie s veľkým divadlom a pýchou, ale ako jednoduché ohlásenie – a to povzbudilo druhých, aby urobili to isté. Tí, ktorí nemali čo dať, dostali ticho; tí, ktorí mali námietky, si na ich vyjadrenie vybrali nejaký skorší alebo neskorší čas. To je pravdepodobne ten spôsob, ako by sa veci *mali* robiť v príčetnom ľudskom spoločenstve – zohľadniť, že ľudia často majú problém motivovať sa tak, ako by chceli byť, a že sa tomu dá pomôcť spoločenským povzbudením prekonať túto slabosť vôle.

Ale aj keď nesúhlasíte s touto časťou, stále si povedzme, že aj súhlasiace aj nesúhlasiace názory by sa mali vyjadriť verejne. Ak podporujúci čelia napohľad pevnej stene námietok a nesúhlasu – aj keby to vyplývalo z ich vlastnej samocenzúry pred nepohodlím – to *nie je* skupinová rozumnosť. Je to iba *zrkadlový obraz* toho, čo robia skupiny Temnej Strany, aby si udržali svojich nasledovníkov. Obrátená hlúposť nie je inteligencia.

*→

318. Tolerujte toleranciu

Jedna z pravdepodobných vlastností niekoho, kto sa rozhodol byť „racionalista“, je nižšia než obvykle tolerancia voči chybám v uvažovaní. Nie je to nevyhnutné. Mohli ste, povedzme, odmietnuť svoje náboženstvo skrátka preto, lebo ste si všimli *viac* dier alebo *hlbšie* diery v uvažovaní, a nie preto, že by vás vďaka vašej povahe *viac hnevala* chyba konštantnej veľkosti. Ale realisticky vzaté, mnohí z nás asi majú úroveň „nahnevanosti na všetky tieto chyby, ktoré som si všimol“ nastavenú vyššie než je priemer.

Preto je také dôležité, aby sme tolerovali toleranciu druhých, ak chceme niečo spolu dosiahnuť.

Pre mňa je vzorovým príkladom tolerancie Ben Goertzel, ktorý okrem iného organizuje každoročnú konferenciu o UI, a kto dokáže povedať niečo pekné *o každom*. Ben sa dokonca pochvalne vyjadril aj o myšlienkach M*nt*f*xa, najslávnejšieho zo všetkých šarlatánov v UI. (Zdá sa, že M*nt*f*x začal pridávať odkaz na Benov kompliment do svojho e-mailového podpisu, asi preto, lebo je to jediný kompliment, ktorý kedy dostal od skutočného akademika UI.) (Prosím *nevyslovujte* jeho Skutočné Meno správne, lebo ho tým privoláte.)

→ http://lesswrong.com/lw/3h/why_our_kind_cant_cooperate/

Ale časom som pochopil, že je to jedna z Benových silných stránok – že je milý k mnohým ľuďom, ktorých by ostatní, vrátane napríklad mňa, ignorovali – a z času na čas sa mu to oplatí.

A ak ja strhnem body z Benovej povesti za to, že dokáže nájsť niečo pekné, čo možno povedať aj o ľuďoch a projektoch, ktoré ja považujem za beznádejné – dokonca aj o *M*nt*f*xovi* – potom vlastne trvám na tom, že Ben musí *neznášať každého, koho neznášam ja*, aby som s ním mohol spolupracovať.

Je toto realistická latka? Najmä ak rôznych ľudí hnevajú rôzne množstvá rozličných vecí?

Ale je ťažké si to pamätať, keď je Ben milý *k toľkým* idiotom.

Spolupráca je nestabilná, v teórii hier aj v evolučnej biológii, bez *nejakého* trestu za podraz. Jedna vec je teda strhnúť niekomu body z jeho povesti za chyby, ktoré urobil *on sám, priamo*. Ale ak na niekoho krivo zazeráte za to, že *odmieta karhať* osobu alebo myšlienku, tak toto je *trestanie netrestajúcich*, omnoho nebezpečnejší princíp, ktorý dokáže zamknúť rovnováhu na danom mieste dokonca aj keď je škodlivá pre *všetkých* prítomných.

Nebezpečenstvo trestania netrestajúcich je niečo, čo si musím pripomínať zakaždým, keď Robin Hanson poukáže na chybu v nejakom akademickom klišé a predsa skromne prizná, že by sa mohol mýliť (a nemýli sa). Alebo vždy, keď vidím ako Michael Vassar stále zvažuje potenciál niekoho, koho som ja odpísal ako beznádejného 30 sekúnd po našom zoznámení. Musím si pripomínať: „Toleruj toleranciu! Nežiadaj, aby tvoji spojenci boli *rovnako extrémni* vo svojom negatívnom hodnotení všetkého, čo sa tebe nepáči!“

Ja *mám* takú povahu, že ma vytočí, keď sa zdá, že je niekto iný príliš zhovievavý. Neviem, či je každý taký, ale mám podozrenie, že prinajmenšom *niektorí* moji kolegovia ašpirujúci racionalisti takí sú. Nebol by som prekvapený, keby sa ukázalo, že je to všeobecný ľudský pocit; má zrejmé evolučné zdôvodnenie – ktoré by z neho robilo veľmi *nepríjemnú* a *nebezpečnú* adaptáciu.

Vo všeobecnosti nie som fanúšikom „tolerancie“. Rozhodne neverím v to, že treba „tolerovať netoleranciu“, ako niektorí nekonzistentne tvrdia. Ale budem sa snažiť tolerovať *ľudí, ktorí sú tolerantnejší než ja*, a súdiť ich iba podľa ich *vlastných* nevypožičaných chýb.

Ach, a netreba dodávať, že ak ľudia zo Skupiny X na vás hľadia s očakávaním, čakajú na to, kedy budete nenávidieť tých správnych nepriateľov s tou správnou intenzitou, a sú pripravení karhať vás, ak vy nebudete karhať dosť nahlas, potom ste si možno našli nesprávnu skupinu.

Akurát nežiadajte, aby *každý*, s kým pracujete, bol rovnako netolerantný voči takémuto správaniu. Odpustite svojim priateľom, ak niektorí z nich naznačia, že Skupina X možno nie je až taká hrozná...

*→

319. Cena za vaše zapojenie sa

V hre Ultimátum si prvý hráč vyberie, ako rozdeliť 10 dolárov medzi seba a druhého hráča, a druhý hráč sa rozhodne, či s rozdelením súhlasí alebo ho odmieta – v tom druhom prípade nikto nedostane nič. Podľa bežnej kauzálnej teórie rozhodovania (dve krabičky v Newcombovom probléme, podraziť vo Väzenskej dileme) by mal druhý hráč uprednostniť ľubovoľné nenulové množstvo pred ničím. Ale ak prvý hráč *očakáva* takéto správanie – prijať ľubovoľnú nenulovú ponuku – potom nemá motív ponúknuť viac ako cent. Predpokladám, že už všetci viete, že nie som fanúšikom bežnej kauzálnej teórie rozhodovania. Tí z nás, ktorí majú stále záujem spolupracovať vo Väzenskej dileme, či už preto, lebo je iterovaná, alebo preto, lebo máme vo svojej funkcii úžitku výraz pre spravodlivosť, alebo preto, lebo používame nezvyčajnú teóriu rozhodovania, môžeme takisto túto ponuku jedného centu neprijať.

A v skutočnosti väčšina „rozdeľovačov“ v Ultimáte ponúkne rovnaký podiel; a väčšina „schvaľovačov“ v Ultimáte odmietne ponuky menšie než 20 %. Hra o 100 amerických dolárov hraná v Indonézii (priemerný ročný prijem na osobu v tej dobe: 670 amerických dolárov) ukázala odmietnuté ponuky 30 dolárov, hoci sa to rovná mzde za dva týždne. Asi môžeme predpokladať, že hráči v Indonézii

→ http://lesswrong.com/lw/42/tolerate_tolerance/

nerozmýšľali nad akademickými debatami o Newcombovom probléme – toto je skrátka spôsob, ako sa mnohí ľudia cítia pri hre Ultimátum, ešte aj keď sa hrá o skutočné peniaze.

Existuje analógia hry Ultimátum pri skupinovej koordinácii. (Študoval ju niekto? Dúfam...) Povedzme, že existuje nejaký spoločný projekt – vlastne povedzme, že je to altruistický spoločný projekt zameraný na pomoc obetiam lúpeží v Kanade, *alebo niečo*. Ak sa pridá pridáte k tomuto skupinovému projektu, urobíte tak viac, než by ste dokázali sám, vzhľadom na vašu funkciu úžitku. Takže je jasné, že by ste sa mali pridať.

Ale moment! Projekt proti lúpežiam drží svoje financie investované na finančnom trhu! To je smiešne; to im nevynesie ani taký útok ako americké štátne dlhopisy, a tobôž nie toľko ako indexový fond vyplácajúci dividendy.

Je jasné, že tento projekt riadia idioti, a nemali by ste sa zapájať, dokiaľ nezmenia svoje spôsoby chybného investovania.

Teraz si možno uvedomíte – ak ste sa zastavili, aby ste nad tým porozmýšľali – že po zvážení všetkých vecí, by ste *stále* urobili lepšie, keby ste spolupracovali s týmto spoločným protilúpežným projektom, než keby ste sa pustili do svojho vlastného boja proti zločinu. Lenže potom – možno si tiež uvedomíte – ak *príliš ľahko odsúhlasíte*, že sa pridáte do skupiny, aký budú mať potom *motív* zmeniť svoje chybné investovanie?

No... Okej, pozrite. Možno preto, lebo už nežijeme v pravekom prostredí, kde každý poznal každého... a možno preto, lebo sa dav nekonformistov snaží zapudiť *normálne* sily skupinovej súdržnosti ako je konformita a uctievanie vedenia...

…zdá sa mi, že ľudia v zhluku ateistov / libertariánov / technofilov / fanúšikov sci-fi / a tak ďalej často nastavujú svoju cenu za zapojenie sa *veľmi veľmi veľmi* vysoko. Ako keby sme hrali Ultimátum s delením medzi 50 hráčov, kde každý jeden z 50 hráčov trvá na tom, že musí dostať aspoň 20 % peňazí.

Ak sa zamyslíte, ako často takéto situácie mohli nastávať v pravekom prostredí, potom je to takmer isto otázkou evolučnej psychológie. Emócie systému 1, nie výpočty systému 2. Naše intuície, kedy sa pripojiť do skupín, verzus kedy sa zdráhať, aby sme dosiahli väčšie ústupky voči nášmu uprednostňovanému spôsobu robenia vecí, by boli dobre vyladené v prostredí lovcov-zberačov, kde bolo napríklad 40 ľudí, ktorých ste všetkých osobne poznali.

A keď sa tá skupina skladá z 1000 ľudí? Potom vaše inštinkty lovca-zberača podcenia zotrvačnosť takejto veľkej skupiny, a budú žiadať nereálne vysokú cenu (v strategickom odklone) za vaše zapojenie. Existuje len obmedzené množstvo organizačného úsilia, a obmedzené množstvo stupňov voľnosti, ktoré možno vynaložiť, aby sa vyhovelo požiadavkám každého človeka.

A ak je stratégia veľká a zložitá, také niečo, čo zaberie napríklad desiatim ľuďom papierovačky na celý týždeň, namiesto zbúchania za polhodinu jednania pri táboráku? Potom vaše inštinkty lovca-zberača podcenia zotrvačnosť skupiny, relaívne k vašim vlastným požiadavkám.

A ak žijete vo svete väčšom než jeden kmeň lovcov-zberačov, takže vidíte iba jedného predstaviteľa skupiny, ktorý jedná s vami, a nie sto ďalších jednaní, ktoré sa už odohrali? Potom vám vaše inštinkty povedia, že je to iba jeden človek, navyše cudzinec, a vy dvaja ste si rovní; hocijaké myšlienky, ktoré k rokovaciemu stolu prináša on, sú rovné hocijakým myšlienkam, ktoré prinášate vy, a mali by ste sa stretnúť zhruba uprostred.

A ak trpíte hocijakou slabosťou vôle alebo akráziou, alebo ak vás ovplyvňujú motívy iné než tie, o ktorých by ste si pripustili, že vás ovplyvňujú, potom ľubovoľný skupinový altruistický projekt, ktorý vám neposkytuje odmeny v postavení a moci, sa môže ľahko ukázať ako nehodný vašej pozornosti.

Tu pripúšťam, že hovorím v prvom rade z pohľadu niekoho, kto sa pokúša pásť stádo mačiek; a nie z druhej strany ako niekto, kto sa väčšinu svojho času snaží uchrániť si svoju energiu, aby mohol vydierať tých prekliatych idiotov, ktorí už sú v projekte. Možno mám trochu predsudky.

Ale zdá sa mi, že rozumné a jednoduché pravidlo by mohlo byť:

Ak, celkovo, pripojenie tvojich síl k skupinovému projektu *by malo stále v čistom pozitívny účinok* podľa vašej funkcie úžitku...

(alebo väčší pozitívny účinok než ľubovoľné iné hraničné využitie, ktorému by ste inak venovali tieto zdroje, hoci tento spôsob uvažovania mi pripadá menej používaný a nerealistický pre človeka, z dôvodov, o ktorých možno napíšem raz neskôr)

...a tá príšerne hrozná otravná vec nie je taká dôležitá, aby ste sa *vy osobne* zapojili dostatočne hlboko a venovali tomu toľko hodín, týždňov alebo rokov, koľko bude treba na jej opravenie...

...potom táto téma nestojí za to, aby ste odopierali svoju energiu tomuto projektu; ani inštinktívne, dokiaľ neuvidíte, že vám ľudia venujú pozornosť a vážia si vás, ani vedomým zámerom vydierať skupinu, aby to urobila.

A ak pre vás táto téma *znamená* tak veľa... potom sa rozhodne pridajte do danej skupiny a urobte všetko, čo treba, aby ste veci napravili.

Jedine, keby vám to existujúci prispievatelia odmietli dovoliť, *a* dalo by sa očakávať, že rozumná tretia strana by došla k záveru, že ste dosť schopný na to, aby ste to urobili, *a* ak tam nie je nikto iný, komu by ste týmto liezli do kapusty, tak *potom*, možno, tu máme problém. A možno je vhodný čas na trochu vydierania, ak prostriedky, ktoré môžete podmienečne prisľúbiť, sú dosť veľké na to, aby im stáli za pozornosť.

Je toto pravidlo trochu extrémne? Ach, možno. Rozhodujúci mechanizmus projektu by *mal* mať motív byť zodpovedný k svojim podporovačom; nepodmienená podpora by vytvárala svoje vlastné problémy.

Ale *zvyčajne*... pozorujem, že ľudia podceňujú cenu toho, čo žiadajú, alebo možno konajú podľa inštinktu a nastavia svoju cenu *veľmi veľmi veľmi* vysoko. Ak chce dav nekonformistov niekedy niečo urobiť spoločne, potrebujeme sa posunúť smerom k pridávaniu sa do skupín a zostávaniu tam aspoň *o trochu* ľahšie. Dokonca aj zoči-voči mrzutostiam a nedokonalostiam! Dokonca aj zoči-voči ignorovaniu našich vlastných lepších nápadov!

V dobe internetu a v spoločnosti nekonformistov začína byť trochu únavné čítať 451. verejný e-mail od niekoho, kto hovorí, že Spoločný Projekt nie je hoden jeho podpory, dokiaľ na webovej stránke nepoužijú bezpätkové písmo.

Samozrejme toto zvyčajne nie je o písme. Môže o byť o lenivosti, akrázii, alebo skrytých odmietnutiach. Ale slovami skupinových noriem... slovami toho, aké verejné vyhlásenia si vážime, a akými výhovorkami verejne pohŕdame... by sme asi *chceli* povzbudiť takúto skupinovú normu:

Ak daná téma nie je hodna toho, aby ste ju osobne opravili, nech to stojí hocikoľko úsilia, a ak nie je výsledkom priamo zlého zámeru, nie je hodno odmietnuť svoju pomoc kauze, ktorú považujete za hodnotnú.

*→

320. Dokáže humanizmus dorovnať výkon náboženstva?

Možno najväčšia *dobrovoľná* organizácia v našom modernom svete – spojená nie políciou a daňami, nie výplatami a menežermi, ale dobrovoľnými príspevkami prúdiacimi od jej členov – je katolícka cirkev.

Je príliš veľká na to, aby držala pokope vyjednávaním medzi jednotlivcami, ako pracovná skupina v tlupe lovcov-zberačov. Ale vo väčšom svete, kde je možné nakaziť viac ľudí a rýchlejšie prenášať infekciu, môžeme očakávať virulentnejšie mémy. Starý Zákon nehovorí o pekle, ale Nový Zákon áno. Katolícku cirkev držia pohromade afektívne špirály smrti – ohľadom myšlienok, inštitúcií, a vodcov. Sľuby večného šťastia a večného zatratenia – teológovia v tieto veci naozaj neveria, ale mnohí bežní katolíci áno. Jednoduchá konformita ľudí, ktorí sa stretávajú osobne v kostole a sú pod tlakom svojho okolia. Atď.

My, ktorí máme tú opovážlivosť hovoriť si „racionalisti“, sa považujeme za príliš dobrých na takéto spoločné putá.

→ http://lesswrong.com/lw/5j/your_price_for_joining/

A tak hocikto, kto má *jednoduchý* a *zrejmý* charitatívny projekt – napríklad poslať potraviny a prístrešky obetiam prílivovej vlny v Thajsku – by dopadol *omnoho* lepšie, keby poprosil pápeža, aby mobilizoval katolíkov, než Richarda Dawkinsa, aby mobilizoval ateistov.

Dokiaľ je toto pravda, ľubovoľné zvýšenie ateizmu na úkor katolicizmu bude trochu prázdnym víťazstvom, bez ohľadu na všetok ostatný úžitok.

Iste, katolícka cirkev zároveň odporuje používaniu kondómov v Afrike pustošenej AIDS. Iste, plytvajú hromadami vyzbieraných peňazí na všetky tie náboženské veci. Dopriavať si nejasné myslenie nie je bez následkov, za modlitbu sa platí.

Zdržať sa robenia škodlivých vecí, to *je* skutočné víťazstvo pre racionalistu...

Pokiaľ to nie je vaše *jediné* víťazstvo, lebo v tom prípade to vyzerá trochu prázdne.

Pokiaľ *odhliadneme od všetkých škôd*, ktoré urobila katolícka cirkev, a pozrieme sa *iba* na to dobré... robí potom priemerný katolík *v hrubom* viac dobra než priemerný ateista, jednoducho tým, že je aktívnejší?

Možno ak ste múdrejší, ale menej motivovaný, hľadáte vysoko efektívne zásahy a kupujete utilony lacno... Ale iba málo z nás to *naozaj* robí, na rozdiel od plánovania, že to jedného dňa urobia.

Teraz by ste mohli rozhodiť rukami a povedať: „Dokiaľ nemáme priamu kontrolu nad motivačnými obvodmi v našich mozgoch, nie je realistické očakávať, že racionalista bude rovnako silno motivovaný ako niekto, kto naozaj verí, že bude večne horieť v pekle, ak neposlúchne."

To je férový bod. Hocijaká ľudová veta v tom zmysle, že rozumný činiteľ by mal fungovať prinajmenšom rovnako dobre ako nerozumný činiteľ, je založená na predpoklade, že ten rozumný činiteľ sa vždy dokáže jednoducho držať hocijakých „nerozumných" pravidiel, o ktorých vidí, že vyhrávajú. Lenže ak si nemôžete *vybrať* neobmedzenú myšlienkovú energiu, potom sa môže stať, že niektoré nepravdivé názory sú, naozaj, omnoho silnejšie motivujúce než hocijaké dostupné pravdivé názory. A ak vo všeobecnosti trpíme altruistickou akráziou, nedokážeme sa primäť pomáhať toľko, koľko si myslíme, že by sme mali, potom je možné, že tí bohabojní vyhrajú v súťaži o altruistický výkon.

Ale aj keď je to motivované pokračovanie, pouvažujme nad touto otázkou o čosi dlhšie.

Dokonca ani strach z pekla nie je dokonalá motivácia. Obojok evolúcie nedáva ľuďom tak veľa voľnosti; dokážeme krátkodobo odolávať pokušeniu, ale potom sa nám minie myšlienková energia. Dokonca ani viera, že pôjdete do pekla, nezmení tento holý fakt o mozgových obvodoch. Takže veriaci hrešia, a potom ich trápi predstava, že pôjdu do pekla, dosť podobne ako si fajčiari sami vyčítajú, že nedokážu prestať fajčiť.

Keby skupine racionalistov na niečom *veľmi* záležalo... kto hovorí, že by nedokázali dorovnať skutočný, de-facto výkon veriacich katolíkov? V stávke nemusí byť „nekonečné" šťastie ani „večné" zatratenie, ale mozog si samozrejme nedokáže predstaviť ani 3^^^3, tobôž nekonečno. Kto hovorí, že skutočné množstvo neurotransmiterov záujmu v mozgu (takpovediac) musí byť omnoho menšie pre „rast a prekvitanie ľudstva" alebo dokonca „prílivom zasiahnutých Thajcov" než pre „večné šťastie v Nebi"? Čokoľvek, čo sa týka viac než 100 ľudí, zahŕňa úžitky príliš veľké na predstavivosť. A vyskytujú sa tu všemožné iné štandardné skreslenia; vedieť o nich by mohlo poskytnúť nejaký bonus, dúfajme?

Kognitívne behaviorálna terapia a zenová meditácia sú dve myšlienkové disciplíny, ktoré experimentálne ukázali, že vedú k skutočným zlepšeniam. Nie je to tá oblasť umenia, na ktorej rozvoj sa sústredím, ale ja predsa nemám za sebou skutočné bojové umenie rozumnosti. Ak spojíte cieľ, o ktorý sa naozaj oplatí starať, s disciplínou získanou z KBT a zenovej meditácie, kto potom hovorí, že racionalisti neudržia krok? Alebo ešte všeobecnejšie: ak máme umenie boja s akráziou založené na dôkazoch, a experimenty, aby sme videli, čo naozaj funguje, potom kto hovorí, že musíme byť menej motivovaní než nejaká nezorganizovaná myseľ, ktorá sa bojí Božieho hnevu?

Ale... to je špekulácia ohľadom ďalekej budúcnosti, že sa možno podarí vyvinúť umenie, ktoré dnes neexistuje. Nie je to technika, ktorú by som mohol použiť hneď teraz. Predkladám ju len, aby som ilustroval myšlienku, že *netreba tak rýchlo lámať palicu nad rozumnosťou*: Pochopiť, čo je nesprávne, pokúsiť sa to inteligentne opraviť, a zbierať indície, či to fungovalo alebo nie – to je silný prístup, ktorého sa netreba zľahka vzdať pri pohľade na prvú prekážku.

Naozaj, myslím si, že o čo tu ide, menej súvisí s motivujúcou silou večného zatratenia, a omnoho viac s motivujúcou silou *fyzického stretávania sa* s inými ľuďmi, ktorí majú spoločný cieľ. Inými slovami, je to sila toho, že ste fyzicky prítomní v kostole a máte veriacich susedov.

Toto je problém pre racionalistické spoločenstvo v jeho terajšom štádiu rastu, pretože je nás málo a sme zemepisne rozdelení po celom širom svete. Keby čitatelia tohto blogu bývali v okruhu 10 kilometrov jeden od druhého, stavím sa, že by sme urobili omnoho viac, nie z dôvodov *koordinácie*, ale len z čírej *motivácie*.

Neskôr napíšem nejaké dlhodobé, vizionárske, idealistické myšlienky ohľadom tohto konkrétneho problému. Lepšie by však boli krátkodobé riešenia, ktoré nezávisia na 100-násobnom zvýšení nášho počtu. Konkrétne rozmýšľam, či by nám najlepšie moderné videokonferenčné programy neboli poskytnúť niečo z motivačného efektu osobného stretnutia s niekým; obávam sa, že odpoveď je „nie“, ale možno sa to oplatí skúsiť.

Medzičasom... z krátkodobého hľadiska sme zaseknutí v boji s akráziou prevažne bez posilňujúcej osobnej prítomnosti iných ľudí, ktoým na tom záleží. Chcem povedať niečo ako: „Je to ťažké, ale *dá* sa to,“ akurát si nie som istý, či je to vôbec pravda.

Mám podozrenie, že *najväčší* krok, aký by racionalisti mohli urobiť smerom k dorovnaniu výkonu per capita katolíckej cirkvi by bolo mať pravidelné fyzické stretnutia ľudí prispievajúcich k tej istej úlohe – nie za účelom koordinácie, iba za účelom motivácie.

A bez toho...

Mohli by sme vyskúšať ako skupinovú normu povoliť – čoby, chváliť – silný záujem o niečo. A skupinovú normu, že sa čaká, že so svojím životom robíte niečo užitočné – prispejete svojím dielom k upratovaniu tohto sveta. Náboženstvo v skutočnosti až tak nezdôrazňuje tú stránku, že treba veci *urobiť*.

A keby racionalisti dokázali dorovnať *polovicu* priemerného altruistického výkonu na katolíka, potom si nemyslím, že je čo len *trochu* nereálne predpokladať, že s lepším zameraním na efektívnejšie kauzy by typický racionalista dokázal urobiť dvakrát viac.

Koľko veľa zo svojich príjmov míňa katolicka cirkev na všetky tie zbytočné náboženské veci namiesto skutočnej pomoci ľuďom? Viac ako 50 %, stavil by som sa. Mohli by sme teda povedať – s istou iróniou, hoci to nie je celkom ten duch, v ktorom by sme mali robiť veci – že by sme sa mali pokúsiť rozšíriť skupinovú normu darovať aspoň 5 % príjmu na *skutočné* kauzy. (10 % je zvyčajne odporúčaný náboženský desiatok.) A potom je tu umenie vyberania si káuz, kde sú očakávané utilony rádovo lacnejšie (dokiaľ pretrváva neefektívny trh utilonov).

Ale dávno predtým, než môžeme začať snívať o takomto chvastaní sa, musíme my svetskí humanisti popracovať na tom, aby sme aspoň *dorovnali* per capita dobročinný výkon veriacich.

* →

321. Cirkev verzus pracovná skupina

Vo všeobecnosti hľadím s podozrením na závisť voči bláznivým skupinám alebo snahu slepo kopírovať rytmus náboženstva – nazýval som to „chválospevy na neexistenciu Boha“ a odpovedal som: „Dobrý ‚ateistický chválospev‘ je jednoducho pieseň o hocičom, o čom sa oplatí spievať, čo zhodou okolností nie je náboženské.“

Lenže náboženstvo napĺňa isté diery v mysliach ľudí, z ktorých niektoré sa dokonca oplatí naplniť. Ak odstránite náboženstvo, musíte si uvedomovať, ktoré medzery vám zostanú.

Keby ste náhle vymazali náboženstvo zo sveta, najväčšia zostávajúca medzera by nebola niečo ohľadom ideálov alebo morálky; bola by to cirkev, spoločenstvo. Spomedzi tých, ktorí dnes zostávajú ako veriaci, hoci v skutočnosti v Boha neveria – koľko z nich zostáva preto, lebo chcú zostať v kontakte so svojimi susedmi v kostole, so svojou rodinou a priateľmi? Koľkí by konvertovali na ateizmus, keby všetci

→ http://lesswrong.com/lw/5t/can_humanism_match_religions_output/

títo ostatní dekonvertovali, a keby *toto* bola cena za zotrvanie v spoločenstve a zachovanie si ich úcty? Tipol by som... že asi mnohí.

V skutočnosti... je to asi niečo, čo som sám celkom dobre nepochopil. „Koláčiky a babysitting," boli prvé dve veci, ktoré mi napadli. Naozaj cirkvi ponúkajú pomocnú ruku v núdzi? Alebo iba plece, na ktorom sa možno vyplakať? Aké silná je cirkevná komunita? Asi to závisí od konkrétnej cirkvi, a v každom prípade to nie je tá správna otázka. Človek by mal začať uvažovaním o tom, čo dáva svojim ľuďom tlupa lovcov-zberačov, a opýtať sa, čo z toho chýba v modernom živote – ak moderná cirkev v rozvinutej krajine plní iba *niečo* z toho, potom by sme sa rozhodne mali pokúsiť urobiť to *lepšie*.

Takže *bez* kopírovania náboženstva – *bez* predpokladania, že sa *musíme* zhromaždiť každú nedeľu ráno v budove s oknami z farebného skla, s formálne oblečenými deťmi, a počúvať ako niekto spieva – zamyslime sa, ako zaplniť túto emocionálnu medzeru, keď náboženstvo prestane byť možnosťou.

Aby sme vykročili zo šablóny – zo zvieracej kazajky uložených myšlienok o tom, ako sa takéto veci robia – uvážte, že *niektoré* moderné kancelárie môžu tiež plniť rovnakú rolu ako cirkev. Čím myslím, že niektorí ľudia majú to šťastie, že nájdu spoločenstvo na svojom pracovisku: priateľských kolegov, ktorí napečú koláčiky pre celú kanceláriu, ktorých puberťákov možno bezpečne najať na babysitting, a možno by dokonca pomohli v prípade katastrofy...? Ale iste nemá každý toľko šťastia, aby našiel spoločenstvo v kancelárii.

Rozmýšľajme ďalej – cirkev je *oficiálne* na uctievanie, a pracovisko je *oficiálne* na obchodné ciele organizácie. Ani jedno z toho nie je starostlivo *optimalizované*, aby slúžilo ako spoločenstvo.

Pri pohľade na typickú náboženskú cirkev, by ste napríklad mohli mať podozrenie – hoci by bolo lepšie všetky tieto veci testovať experimentálne, než iba podozrievať...

- Že vstávanie v nedeľu ráno nie je optimálne;
- Že nosenie formálneho oblečenia nie je optimálne, najmä pre deti;
- Že počúvanie, ako tá istá osoba káže každý víkend na tú istú tému („náboženstvo") nie je optimálne;
- Že výdavky na cirkev a kňaza sú vysoké, v porovnaní s počtom rôznych spoločenstiev, ktoré by sa mohli v rôznych časoch deliť o tú istú budovu pri stretávaní sa;
- Že pravdepodobne neslúžia dosť dobre ani ako zoznamka, pretože cirkvi si myslia, že musia presadzovať svoju stredovekú morálku;
- Že by celá táto vec mala byť témou experimentálneho zbierania údajov, aby sa zistilo, čo funguje a čo nie.

Horeuvedeným používaním slova „optimálne" myslím „optimálne podľa kritérií, ktoré by ste použili, keby ste vyslovene budovali spoločenstvo *ako* spoločenstvo." Minúť kopec peňazí na vyčačkaný kostol s oknami z farebného skla a kňazom na plný úväzok dáva zmysel iba ak naozaj *chcete* míňať peniaze na náboženstvo *ako* náboženstvo.

Priznávam sa, že keď kráčam vedľa kostolov v mojom meste, moja hlavná myšlienka je: „Tieto budovy vyzerajú naozaj, naozaj draho, a je ich príliš veľa." Keby ste to celé robili od nuly... potom by ste možno mali veľkú budovu, ktorú možno používať na občasné svadby, ale bola by rozdelená pre rôzne spoločenstvá, ktoré by sa stretávali v rôznom čase cez víkend, a mala by aj peknú veľkú obrazovku, ktorá by sa dala použiť, aby rečníci ukazovali prezentácie, lektori niečo učili, alebo by niekto mohol pozerať filmy. Farebné sklo? To nie je vysoká priorita.

Alebo miera, do akej členstvo v cirkvi poskytuje pomocnú ruku v časoch problémov – dalo by sa to zlepšiť osobitným fondom na zlé časy, alebo zmluvou s poisťovňou, keď si uvedomíte, že je to dôležitá funkcia? Možno *nie*; zaťahovať explicitné financie do iných vecí čudne mení ich povahu. Možno práve naopak, dodržiavať isté pravidlá poistenia by mohlo byť *podmienkou* členstva, aby ste sa na spoločenstvo *príliš* nespoliehali... Ale opäť, do tej miery, ako cirkvi poskytujú spoločenstvo, skúšajú to urobiť bez toho, aby naozaj *priznali*, že toto je prakticky všetko, čo z toho ľudia budú mať. To isté platí pre firmy, ktorých pracoviská sú dosť priateľské na to, aby slúžili ako spoločenstvá; je to stále tak trochu náhodná funkcia.

Keď raz začnete rozmýšľať *explicitne*, ako dať ľuďom tlupu lovcov-zberačov, do ktorej môžu patriť, začnete vidieť všelijaké veci, ktoré vyzerajú ako dobré nápady. Mali by ste medzi sebou vítať

nových prisťahovalcov? Kňaz vám o tom môže niekedy urobiť kázeň, ak si myslíte, že cirkev je o náboženstve. Ale ak je vaším vysloveným cieľom budovať spoločenstvo – potom hneď po presťahovaní je čas, keď človeku spoločenstvo najviac chýba, keď najviac potrebuje vašu pomoc. Je to tiež pre tlupu príležitosť rásť. Ak niečo, tlupy by mali súťažiť so štvrťročnými vyhodnotením, kto zachytí viac nových prisťahovalcov.

Ale môžete mať naozaj spoločenstvo, ktoré je *iba* spoločenstvo – ktoré zároveň nie je kancelária alebo náboženstvo? Spoločenstvo, ktoré nemá iný cieľ okrem seba samého?

Možno *áno*. Napokon, majú kmene lovcov-zberačov nejaký cieľ okrem seba samých? - no, išlo tam o prežitie a najedenie sa, to bol cieľ.

Ale hocičo, čo majú ľudia spoločné, najmä hocijaký *cieľ*, ktorý majú spoločný, má sklon *chcieť* si definovať spoločenstvo. Prečo to nevyužiť?

Hoci je toto doba internetu, žiaľ, príliš mnoho spoločných záujmov má priaznivcov príliš široko rozdelených, než aby sformovali slušnú tlupu – ak ste jediným členom cirkvi subgénia vo vašom meste, veľmi vám to asi nepomôže. Naozaj je to iné bez fyzickej prítomnosti; internet *nevyzerá* ako prijateľná náhrada pri súčasnom stave technológie.

Preskočme teda priamo k pointe...

Ak Zem vydrží tak dlho, rád by som videl ako formu racionalistických spoločenstiev pracovné skupiny zamerané na všetku tú prácu, ktorú treba urobiť pri oprave tohto sveta. Spoločenstvá v hocijakej zemepisnej oblasti by sa vytvorili okolo toho najkonkrétnejšieho zhluku, ktorý dokáže uniesť slušne veľkú tlupu. Ak vaše mesto nemá dosť ľudí na to, aby ste si tam našli 50 spoluprogramátorov Linuxu, možno sa budete musieť zmieriť s 15 programátormi open source... alebo v časoch, keď toto celé ešte len začína, s 15 racionalistami, ktorí sa snažia upraviť Zem každý iným spôsobom.

To je to, čo si myslím, že by bolo správnym nasmerovaním energie spoločenstiev a spoločným cieľom, ktorý by ich spájal. Na takéto úlohy tak či tak potrebujeme spoločenstvá a na tejto Zemi je kopec práce, ktorú treba urobiť, takže nemá zmysel plytvať. Treba toho urobiť tak veľa – nech si energia, ktorá sa kedysi vyhadzovala do prázdnoty náboženských inštitúcií, nájde východisko tam. A nech ciele, ktoré možno obdivovať bez nutnosti klamania, zaplnia prázdnotu v štruktúre spoločenstva, ktorú tam zanechá vymazanie náboženstva a jeho fiktívnych vyšších cieľov.

Silné spoločenstvá vytvorené okolo hodnotných cieľov: V takomto tvare by som rád videl dobu po náboženstve, alebo hocijaký zlomok ľudstva, ktorý sa vo svojom živote dostane tak ďaleko.

Hoci... pokiaľ tak či tak máte budovu s peknou veľkou obrazovkou s vysokým rozlíšením, nevadilo by mi skúsiť spochybniť myšlienku, že všetko vzdelávanie po dosiahnutí dospelosti sa musí odohrávať vo vzdialených drahých priestoroch univerzít s učiteľmi, ktorí by radšej robili niečo iné. A je empiricky pravda, že na univerzitách sa spoločenstvám celkom darí. Takže, aby sme boli féroví, sú aj iné možnosti, okolo ktorých by sa dali vybudovať post-náboženské spoločenstvá.

Je toto všetko iba sen? Možno. Pravdepodobne. Nechýba tomu možnosť postupného zavedenia, ak máte dostatok racionalistov v dostatočne veľkom meste, ktorí o tejto myšlienke počuli. Ale len pre ten prípad, že by sa rozumnosť úspešne rozšírila, alebo že by Zem vydržala tak dlho, a že by bolo počuť môj hlas, tak toto je smer, ktorým by som rád videl, aby sa veci hýbali – ako upadajú cirkvi, nepotrebujeme umelé kostoly, ale potrebujeme nové vzory spoločenstva.

⁂→

322. Rozumnosť: spoločný záujem mnohých záujmov

Stránka LessWrong sa príliš netají tým, že existujú mnohé záujmy, ktorým by rozšírenie rozumnosti prospelo – pretože treba trochu viac rozumnosti než zvyčajne na pochopenie ich prípadu, ako podporovateľ alebo aspoň ako prívetivý pozorovateľ. Nie iba samozrejmé prípady ako ateizmus, ale veci ako legalizácia marihuany – kde by ste si želali, aby si ľudia o trochu viac uvedomovali svoje vlastné

→ http://lesswrong.com/lw/5v/church_vs_taskforce/

motivy a podstatu signalizácie, a nechali sa trochu viac ovplyvniť nepohodlnými chladnými faktmi. Výskumný ústav strojovej inteligencie je iba nezvyčajne extrémnym prípadom tohto, kde to zašlo až do toho bodu, že po rokoch márneho úsilia som skrátka mávol rukou a pustil som sa explicitne do úlohy vytvoriť racionalistov.

Ale samozrejme, nie *všetci* racionalisti, ktorých vytvorím, budú mať záujem o *môj* projekt – *a to je v poriadku*. Nemôžete zachytiť *všetku* hodnotu, ktorú vytvoríte, a také úsilie môže mať zlé vedľajšie účinky.

Keby boli podporovatelia iných záujmov dostatočne osvietení, aby uvažovali podobne...

Potom by všetky záujmy, ktorým by prospelo rozšírenie rozumnosti, mohli mať azda niečo ako štandardizovaný materiál, na ktorý by odkázali svojich podporovateľov – spoločná úloha, centralizovaná aby sa šetrilo námahou – a myslieť na seba ako na tých, čo okolo seba šíria trochu rozumnosti. Nezachytia všetku hodnotu, ktorú vytvoria. A to je v poriadku. Zachytia niečo z hodnoty, ktorú vytvoria druhí. Ateizmus má priamo málo spoločného s legalizáciou marihuany, ale ak aj ateisti aj anti-prohibicionisti sú ochotní ústúpiť kúsok dozadu a povedať niečo o všeobecnom abstraktnom princípe vyrovnávania sa s nepríjemnou pravdou, ktorá sa stavia do cesty peknej oprávnenej tiráde, potom by aj ateizmus aj legalizácia marihuany mali trochu úžitku zo snahy ich oboch.

Ale to si vyžaduje – ja viem, že sa tu opakujem, ale je to dôležité – byť ochotní nezachytiť všetku hodnotu, ktorú vytvoríte. Vyžaduje si to, aby ste si v procese hovorenia o rozumnosti uchovali schopnosť na chvíľu *sklapnúť* ohľadom svojej vlastnej kauzy, napriek tomu, že je to tá najlepšia kauza na svete. Vyžaduje to, aby ste nevnímali tieto ostatné kauzy, a aby oni nevnímali vás, ako konkurenta zápasiaceho o obmedzenú zásobu racionalistov, ktorá môže dať obmedzené množstvo podpory; ale skôr ako vytváranie nových racionalistov a zvyšovanie ich schopnosti podpory. Zožnete iba časť svojho úsilia, ale takisto zožnete aj časť úsilia druhých.

Ak sa vy a oni nedohodnete na všetkom – najmä na prioritách – musíte byť ochotní *sklapnúť* ohľadom nezhody. (Azda okrem špecializovaných miest, mimo debaty verejnosti, kde sa budú explicitne riešiť takéto nezhody.)

Istý človek, ktorý nastupoval ako predseda istej organizácie, raz ukázal na to, že tejto organizácii sa príliš nedarilo s jej heslom: „Toto je *najlepšia* vec, ktorú môžete urobiť“, v porovnaní napríklad s ohromným úspechom Nadácie X-Prize hovoriacej bohatým jednotlivcov: „Toto je *cool* vec, ktorú môžete urobiť.“

Toto je jeden z tých vhľadov, kde neveriacky žmurknete a potom pochopíte, aký veľký to má zmysel. Ľudský mozog nedokáže pochopiť veľké záujmy, a ľudia sa ani zďaleka napodobajú na maximalizátorov očakávaného úžitku, a vo všeobecnosti sme altruistickí akratici. Povedať: „Toto je *najlepšia* vec“ nedá viac motivácie než: „Toto je cool vec.“ Akurát si to vyžaduje omnoho vyššie bremeno dôkazu. A provokuje motiváciu odčerpávajúce porovnania so všetkými inými dobrými vecami, ktoré poznáte (možno tým, že zmenšuje morálne uspokojenie, ktoré ste si už kúpili).

Keby sme pracovali podľa predpokladu, že každý je štandardne altruistický akratik (niekto, kto by chcel, aby sa dokázal rozhodnúť urobiť viac) – ale prinajmenšom, že väčšina možných zaujímavých podporovateľov zodpovedá tomuto popisu – potom bojovanie o to, ktorá úloha si *najviac* zaslúži podporu, by malo za účinok zníženie celkovej zásoby altruizmu.

„Lenže,“ poviete, „doláre sú zameniteľné; dolár, ktorý použijete na jednu vec, nemožno použiť na nič iné!“ Na to odpoviem: Ale ľudia *v skutočnosti nie sú* maximalizátori očakávaného úžitku, ako kognitívne systémy. Doláre pochádzajú z rôznych myšlienkových účtov, stoja rôzne množstvá *sily vôle* (skutočne obmedzujúceho zdroja) za rôznych okolností, ľudia chcú svoje dary distribuovať, ako akt myšlienkového účtovníctva na minimalizáciu ľútosti, keby jedna kauza nevyšla, a povedať niekomu o ďalšej kauze môže zvýšiť celkové množstvo, ktorým pomôže.

Existujú samozrejme hranice tohto princípu dobrotivej tolerancie. Ak má niekto projekt na pomáhanie zatúlaným šteniatkam zostať sa do teplých domovov, pravdepodobne je najlepšie vnímať ho ako niekoho, kto sa pokúša zneužívať chyby v ľudskej psychológii na svoj osobný úžitok, než ako hodnotnú podúlohu veľkého projektu ľudského pokroku Nového Osvietenstva.

Ale do tej miery, do akej niečo naozaj je úlohou, ktorú by ste chceli, aby ľudstvo urobilo... potom individuálne porovnanie tohto projektu s Vaším Obľúbeným Projektom asi nepomôže vášmu projektu tak, ako si myslíte. Možno sa potrebujeme naučiť hovoriť, zo zvyku a takmer v každej diskusii: „Toto je *cool* racionalistický projekt“ a nie „Jedine môj projekt *dáva najviac očakávaných utilonov na hraničný dolár*.“ Ak sa niekto dostatočne chladokrvný na to, aby maximalizoval očakávaný úžitok zameniteľných peňazí bez ohľad na emocionálne vedľajšie účinky, *vyslovene opýta*, možno by sme ho mohli nasmerovať na *špecializované* podfórum, kde si to medzi sebou vybavujú tí, ktorí sú ochotní tvrdiť, že majú nárok na *najvyššiu* prioritu. Hoci ak všetko pôjde dobre, potom projekty, ktoré majú silný nárok tvrdiť, že sú podfinancované, dostanú viac investícií a ich hraničné výnosy klesnú, a víťaz súťažiacich nárokov opäť nebude jasný.

Ak existujú mnohé racionalistické projekty, ktorým by prospelo, keby sa zvýšila hladina príčetnosti, potom si viem predstaviť, že by ich vzájomná tolerancia a spoločná investícia do šírenia rozumnosti mohli vykazovať problém spoločného vlastníctva. Ale nepripadá mi, že by bolo ťažké poradiť si s tým: ak nejaká skupina nie je ochotná rozdeliť sa o racionalistov, ktorých vytvoria, alebo spomenúť, že možno existujú aj iné projekty Nového Osvietenstva, potom ľubovoľné spoločné racionalistické zdroje môžu odstrániť odkaz na ich projekt zo zoznamu cool vecí.

Hoci toto všetko sú idealistické myšlienky zamerané na budúcnosť, úžitok – pre nás všetkých – by mohol byť nájsť pár zaujímavých vecí, ktoré si práve neuvedomujeme. Tak veľa racionalistických projektov má málo podporovateľov, ktorí sú príliš ďaleko; keby sme sa všetci dokázali identifikovať ako prvky Spoločného projektu ľudského pokroku, Nového Osvietenstva, potom by bola podstatne vyššia pravdepodobnosť, že sa desať z nás nájde v ľubovoľnom meste. Práve teraz sú mnohé tieto projekty pre svojich podporovateľov trochu osamelé. Rozumnosť možno nie je *tá najdôležitejšia vec na svete* – tou je samozrejme tá vec, ktorú chránime – ale je to *cool* vec, ktorú máme mnohí spoločnú. Mohli by sme tým získať, ak sa budeme zároveň identifikovať ako racionalisti.

*→

323. Bezmocní jednotlivci

Keď zvážite, že naše skupinové inštinkty sú optimalizované na 50-členné tlupy lovcov-zberačov, kde každý každého pozná, začne to vyzerať ako zázrak, že naše moderné veľké inštitúcie vôbec prežívajú.

Teda – existujú vlády so špecializovanou armádou a políciou, ktoré dokážu vyberať dane. To je nie-praveký mechanizmus, ktorý pochádza z čias vynálezu usadeného poľnohospodárstva a vyberateľných prebytkov; ľudstvo stále zápasí, ako si s tým poradiť.

Existujú korporácie, kde tok peňazí riadi centralizovaný menežment, čo je nie-praveký mechanizmus, ktorý pochádza z čias vynálezu obchodovania vo veľkom s profesionálnej špecializácie.

A vo svete s veľkými populáciami a blízkym kontaktom sa vyvíjajú omnoho nákazlivejšie mémy než v priemernom prípade v pravekom prostredí; mémy, ktoré sa vyhrážajú zatratením, sľubujú nebo, a šíri ich profesionálna trieda kňazov.

Ale vo všeobecnosti odpoveď na otázku: „Ako veľké organizácie prežijú?“ znie: „Nijako!“ Prevažná väčšina veľkých moderných inštitúcií – niektoré z nich extrémne životne dôležité pre fungovanie našej zložitej civilizácie – *v prvom rade ani nikdy neexistovala*.

Po prvýkrát som si to uvedomil, keď som pochopil, ako je financovaná Veda: konkrétne, *nie* darmi jednotlivcov.

Veda je tradične financované vládami, korporáciami, a veľkými nadáciami. Mal som príležitosť osobne odhaliť, aké *úžasne* ťažké je vyzbierať na Vedu peniaze od jednotlivcov. Nie dokiaľ to nie je veda o chorobe s príšernými obeťami, a možno ani vtedy nie.

→ http://lesswrong.com/lw/66/advance_notes_on_the_greater_community/

Prečo? Ľudia sú v skutočnosti prosociálni; dávajú peniaze povedzme na útulky pre šteniatka. Veda je jeden z veľkých spoločenských záujmov, a ľudia si to dokonca všeobecne uvedomujú – prečo teda nie Veda?

Ľubovoľný *konkrétny* vedecký projekt – povedzme štúdovanie genetiky trypanotolerancie u dobytka – sa *emocionálne* dobre nehodí na individuálnu charitu. Veda má dlhý časový horizont, ktorý si vyžaduje dlhodobú podporu. Priebežné alebo dokonca aj záverečné tlačové správy nemusia znieť vôbec emocionálne povzbudzujúco. Nemôžete sa zapojiť ako dobrovoľník; je to práca pre špecialistov. Ak vám ukážeme obrázok vedca, ktorého podporujete priemernou alebo mierne podpriemernou mzdou, nemá to rovnaká dopad, ako keď vám ukážeme šteniatko so širokými očakmi, ktoré ste práve pomohli dopraviť do jeho nového domova. Nedostanete okamžitú spätnú väzbu a pocit okamžitého výkonu, ktorý treba na to, aby jednotlivec ďalej *míňal svoje peniaze*.

Irónia je, že som si to konečne uvedomil nie pri svojej vlastnej práci, ale pri myšlienke: „Prečo čitatelia Setha Robertsa spoločne nepodporia experimentálne testy Robertsovej hypotézy o obezite? Prečo jednotliví filantropi nezaplatia za test Bussardovho polywellového fúzora?“ Toto sú príklady *zrejme* absurdne podfinancovanej vedy, ktorej použitie (keby sa to ukázalo ako pravda) by sa mohli týkať mnohých, mnohých jednotlivcov. Až vtedy mi napadlo, že Veda vo všeobecnosti nie je emocionálne vhodná na to, aby na ňu ľudia míňali vlastné peniaze.

V skutočnosti je na to vhodných *veľmi málo vecí*, pri jednotlivcoch, akých máme teraz. Zdá sa mi, že toto je kľúčom k porozumeniu, prečo svet funguje tak, ako funguje – prečo sa toľko individuálnych záujmov tak málo chráni – prečo má napríklad 200 miliónov dospelých Američanov taký ohromný problém dozerať na 535 členov kongresu.

Ako je teda Veda naozaj financovaná? Vládami, ktoré si myslia, že by mali minúť časť peňazí na Vedu, kde sa tak zákonodarná alebo výkonná moc rozhodne – nie sú to predsa ich *vlastné* peniaze, ktoré miňajú. Dostatočne veľké korporácie sa rozhodnú nečakane hodiť nejakú sumu peňazí na výskum a vývoj. Veľké mimovládne organizácie postavené okolo afektívnych špirál smrti sa môžu pozrieť na vedu, ktorá vyhovuje ich ideálom. Veľké súkromné nadácie, založene na peniazoch, ktoré bohatí jednotlivci vyhradili na svoju povesť, míňajú peniaze na Vedu, ktorá vyzerá, že bude znieť veľmi charitatívne, niečo ako keď sa prideľujú peniaze na orchestre alebo moderné umenie. A potom jednotliví vedci (alebo jednotlivé pracovné skupiny vedcov) navzájom bojujú o ovládnutie tejto vyhradenej zásoby peňazí, ktorá je v rukách členov grantových komisií, ktorí vyzerajú ako ten typ ľudí, ktorí by mal posudzovať vedcov.

Zriedkavo vidíte, že by nejaký vedecký projekt *priamo* žiadal o nejakú časť toku prostriedkov spoločnosti; namiesto toho sa najprv pridelia Vede, a potom vedci bojujú o to, kto ich naozaj dostane. Dokonca aj výnimky z tohto pravidla skôr poháňajú politici (let na Mesiac) alebo vojenské účely (projekt Manhattan) než apel vedcov na verejnosť.

Som si však istý, že keby mala široká verejnosť vo zvyku financovať konkrétnu vedu individuálnymi darmi, celá hromada peňazí by sa vyplytvali napríklad na kvantové bláboly – vychádzajúc z predpokladu, že by široká verejnosť nejako získala zvyk financovať vedu, ale nezmenili by sa žiadne iné fakty o ľuďoch ani o spoločnosti.

Ale stále je zaujímavým bodom, že sa Vede darí prežívať nie preto, lebo je v našom kolektívnom záujme jednotlivcov, aby sa Veda robila, ale skôr preto, lebo sa Veda prisala ako parazit na nové formy veľkých organizácií, ktoré môžu existovať v našom svete. Existuje mnoho iných projektov, ktoré v prvom rade jednoducho nikdy neexistovali.

Zdá sa mi, že sa modernému ľudstvu darí vynakladať pomerne málo koordinovaného úsilia slúžiť kolektívnym záujmom jednotlivcov. Je to skrátka príliš nie-praveký problém, keď máte viac než 50 ľudí. Existujú iba veľkí vyberači daní, veľkí obchodníci, supermémy, občas jednotlivci s veľkou mocou; a pár ďalších organizácii, ako je Veda, ktoré sa na nich môžu paraziticky prisať.

*→

→ http://lesswrong.com/lw/64/helpless_individuals/

324. Peniaze: jednotka záujmu

Steve Omohundro navrhol ľudovú teóriu v tom zmysle, že pre ľubovoľného približne rozumného sebamodifikujúceho činiteľa by hraničný úžitok investovania dodatočných prostriedkov do hocičoho mal byť približne rovnaký. Alebo, aby sme to povedali o trochu exaktnejšie, presunutie jednej jednotky zdrojov medzi ľubovoľnými dvoma úlohami by nemalo spôsobiť zvýšenie očakávaného úžitku vzhľadom na funkciu úžitku daného činiteľa a jeho pravdepodobnostné očakávania ohľadom svojich vlastných algoritmov.

Tento princíp rovnováhy zdrojov naznačuje, že – u veľmi širokého rozsahu približne rozumných systémov, vrátane sebazdokonaľujúcej mysle – bude existovať nejaká spoločná mena očakávaných utilonov, pomocou ktorej možno merať všetko, čo sa oplatí robiť.

V našej spoločnosti sa táto spoločná mena očakávaných utilonov nazýva „peniaze“. Je to miera toho, nakoľko spoločnosti na niečom záleží.

Toto je surový avšak zrejmý bod, ktorý majú mnohí motiváciu popierať.

Pri tomto obecenstve dúfam, že to môžem jednoducho vysloviť a ísť ďalej. Nemysleli ste si predsa doteraz, že „spoločnosť“ je inteligentná, dobrotivá a príčetná, však nie?

Hovorím to, aby som vyslovil určitú pointu, ktorú majú mnohé dobré kauzy spoločnú. Ľubovoľná dobročinná organizácia, o ktorej ste kedy hovorili v dobrom, si iste *želá*, aby ste si túto pointu viac uvedomovali, či už ju niekedy vyslovíte nahlas alebo nie. Počúval som totiž druhých vo svete neziskoviek, a viem, že tu nehovorím iba za seba...

Mnohí ľudia, keď vidia niečo, čo si myslia, že sa oplatí robiť, by radi dobrovoľne venovali pár hodín voľného času, alebo možno päť rokov starý laptop a nejaké konzervy, alebo by niekde pochodovali, ale rozhodne by nemíňali *peniaze*.

Verte mi, rozumiem tomu pocitu. Vždy, keď míňam peniaze, mám pocit, akoby som strácal životy. To je ten problém, keď máte jednotné číslo, ktoré opisuje vašu celkovú hodnotu: Vidieť toto číslo klesať nie je príjemný pocit, napriek tomu, že sa počas vášho bežného života musí rôzne meniť. Mal by proti tomu existovať nejaký princíp teórie zábavy.

Dobre, ale...

V ekonomike *existuje* stará, veľmi stará záhada/pozorovanie, že právnik strávi hodinu ako dobrovoľník v kuchyni pre chudobných, namiesto toho, aby pracoval o hodinu dlhšie a daroval tieto peniaze na zamestnanie niekoho, kto bude v tejto kuchyni pre chudobných pracovať päť hodín.

Existuje niečo, čo sa volá „Ricardov zákon komparatívnych výhod“. Existuje myšlienka zvaná „profesionálna špecializácia“. Existuje pojem „ekonomika rozsahu“. Existuje predstava „vzájomne výhodného obchodu“. Celý dôvod, prečo máme peniaze, je aby sme uskutočňovali *ohromné* výhody možné vďaka tomu, že každý z nás robí to, v čom je *najlepší*.

Toto robia dospelí. Toto je to, čo robíte, keď chcete, aby sa niečo naozaj *urobilo*. Používate *peniaze*, aby ste zamestnali *špecialistov na plný úväzok*.

Áno, ľudia majú občas obmedzenú schopnosť vymieňať svoj čas za peniaze (nezamestnanosť), takže je lepšie, ak môžu darovať priamo to, čo by inak vymenili za peniaze. *Keby* kuchyňa pre chudobných *potrebovala* právnika, a daný právnik by daroval *veľký súvislý* blok právnickej práce *s vysokou prioritou*, *takáto* dobrovoľnícka pomoc by mala zmysel – je to tá istá *špecializovaná* schopnosť, ktorú ten právnik bežne vymieňa za peniaze. Ale „dobrovoľníčenie“ iba jednou hodinou právnickej práce, stále odkladanou, rozloženou na tri týždne cez náhodné minúty medzi inými prácami? To nie je spôsob, akým sa robia veci, *na ktorých niekomu naozaj záleží*, alebo takmer ekvivalentne povedané, *keď ide o peniaze*.

Do tej miery, do akej jednotlivci nedokážu pochopiť tento princíp *na inštinktívnej úrovni*, môžu si myslieť, že používanie peňazí je v istom zmysle *nepovinné* pri dosahovaní vecí, na ktorých záleží iba v *morálnom* zmysle – na rozdiel od úloh ako najesť sa, s ktorých dôležitosťou zaobchádzame veľmi odlišne. Tento faktor možno *samotný* stačí na to, aby nám zabránil dosahovať naše kolektívne spoločné záujmy v skupinách väčších než 40 ľudí.

Ekonomika obchodu a profesionálna špecializácia nie sú iba nejasne dobré a predsa neprirodzene znejúce nápady, *je to jediný spôsob, ako sa kedy čo v tomto svete urobilo*. Peniaze nie sú kúsky papiera, je to *spoločná mena záujmu*.

Preto staré porekadlo: „Peniaze hýbu svetom, láska ho sotva chráni pred výbuchom.“

No a my máme problém s akráziou – neschopnosťou urobiť to, čo sme sa rozhodli urobiť – čo je časť umenia rozumnosti, o ktorej dúfam, že ju vyvinie niekto iný; ja sa skôr špecializujem na oblasť neriešiteľných otázok. A áno, míňať peniaze bolí viac než robiť dobrovoľnícku prácu, pretože môžete vidieť, ako číslo na vašom bankovom účte klesá, zatiaľ čo zostávajúce hodiny našich životov nie sú viditeľne očíslované. Ale keď príde čas najesť sa, pomyslíte si: „Hm, možno by som mohol skúsiť chovať svoj vlastný dobytok, lebo to by ma bolelo menej než míňať peniaze na obed?“ Nie všetko sa dá urobiť *bez* použitia Ricardovho zákona; a na druhej strane tohto obchodu sú ľudia, ktorí cítia rovnakú bolesť pri predstave, že budú mať menej peňazí.

Zdá sa mi takto narýchlo, že by mali existovať veci, ktoré možno urobiť, aby sa zmenšila bolesť zo straty životov, a aby sa zvýšila vnímaná sila súvislosti medzi darovaním peňazí a „urobil som dobrú vec!“ Niečo z toho sa snažím dosiahnuť práve teraz, zdôrazňovaním skutočnej podstaty a moci peňazí; a napádaním jedovatého mému hovoriaceho, že niekto, kto iba dáva peniaze, iste nemá dosť záujmu na to, aby sa osobne zaangažoval. Toto je púhy odraz mysle, ktorá nerozumie pojmu trhovej ekonomiky v dobe po lovcoch a zberačoch. Akt darovania peňazí, to nie je momentálny akt podpísania šeku, je to akt každej hodiny, ktorú ste strávili, aby ste zarobili tie peniaze, aby ste mohli napísať ten šek – rovnako ako keby ste pracovali pre charitu osobne *so svojou profesionálnou kapacitou*, s maximálnou efektivitou dospelých.

Ak ten právnik potrebuje pracovať hodinu v kuchyni pre chudobných, aby si udržal motiváciu a pripomínal si, prečo robí to, čo robí, *to je fajn*. Ale mal by *zároveň* darovať niektoré hodiny, ktoré pracoval v kancelárii, pretože to je moc profesionálnej špecializácie, a takto naozaj robia veci dospelí. Mohol by si predstaviť tento šek ako nákup práva byť dobrovoľníkom v kuchyni pre chudobných, alebo ako potvrdenie času stráveného v kuchyni. O tomto možno neskôr napíšem viac.

Ako prvé priblíženie, peniaze sú jednotka záujmu, až na pozitívny skalárny činiteľ – jednotka relatívneho záujmu. Niektorí ľudia sú šetrní a míňajú menej peňazí *na všetko*; ale ak by ste v skutočnosti minuli 5 dolárov na burrito, potom na čomkoľvek, na čo neminiete 5 dolárov, vám záleží *menej než* na burrite. Ak neminiete dvojmesačnú výplatu na diamantový prsteň, neznamená to, že nemilujete svoju polovičku. („De Beers: Je to iba kameň.“) Ale na druhej strane, ak sa *vždy* zdráhate minúť akékoľvek peniaze na svoju polovičku, a napriek tomu nemáte emocionálny problém minúť 1000 dolárov na plochú obrazovku, potom áno, toto niečo *hovorí* o vašich relatívnych hodnotách.

Áno, šetrnosť je cnosť. Áno, míňanie peňazí bolí. Ale v konečnom dôsledku, ak nie ste nikdy ochotní minúť žiadne jednotky záujmu, znamená to, že nemáte záujem.

* →

325. Kupujte hrejivé pocity a utilony osobitne

V predchádzajúcej kapitole:

> V ekonomike *existuje* stará, veľmi stará záhada/pozorovanie, že právnik strávi hodinu ako dobrovoľník v kuchyni pre chudobných, namiesto toho, aby pracoval o hodinu dlhšie a daroval tieto peniaze na zamestnanie niekoho...
>
> Ak ten právnik potrebuje pracovať hodinu v kuchyni pre chudobných, aby si udržal motiváciu a pripomínal si, prečo robí to, čo robí, *to je fajn*. Ale mal by *zároveň* darovať niektoré hodiny, ktoré pracoval v kancelárii, pretože to je moc profesionálnej špecializácie, a takto naozaj robia veci dospelí. Mohol by si predstaviť tento šek ako nákup práva byť dobrovoľníkom v kuchyni pre chudobných, alebo ako potvrdenie času stráveného v kuchyni.

→ http://lesswrong.com/lw/65/money_the_unit_of_caring/

Keď idú malé staré panie, podržím im otvorené dvere. V skutočnosti si nepamätám, kedy naposledy sa doslova toto stalo (ale som si istý, že to bolo zhruba počas posledného roka). Ale počas posledného mesiaca som bol na prechádzke a našiel som na ceste zaparkované kombi s kufrom celkom otvoreným, takže vnútro auta bolo celkom dostupné. Pozeral som sa, či niekto nevyberá batožiny, ale nebolo to tak. Obzrel som sa, či vôbec niekto niečo robí s týmto autom. A nakoniec som prišiel k domu a zaklopal, potom zazvonil. A áno, ten kufor zabudli otvorený náhodou.

Za iných okolností by toto bol jednoduchý čin altruizmu, ktorý by mohol znamenať skutočný záujem o blahobyt druhých, alebo strach z pocitu viny za nekonanie, alebo túžbu signalizovať dôveryhodnosť sebe alebo druhým, alebo potešenie z altruizmu. Myslím si, že toto všetko sú dokonale legitímne motívy; možno by som dal body navyše za to prvé, ale nestrhával by som žiadne body za to zvyšné. Dôležité je, aby sa ľuďom pomáhalo.

Lenže v mojom vlastnom prípade, keďže ja už pracujem v neziskovom sektore, vzniká ďalšia otázka, či som mohol využiť tých istých šesťdesiat sekúnd *špecializovanejším* spôsobom, aby som druhým priniesol viac úžitku. To jest: či môžem toto naozaj obhájiť ako *najlepšie* využitie môjho času, vzhľadom na ostatné veci, o ktorých tvrdím, že im verím?

Samozrejmá obrana – alebo možno, samozrejmá racionalizácia – je, že takýto čin altruizmu funguje ako obnovovač sily vôle omnoho účinnejšie než povedzme počúvanie hudby. Takisto nedôverujem svojej schopnosti byť altruistom *iba* v teórii; mám podozrenie, že keby som problémy obchádzal, môj altruizmus začne blednúť. Nikdy som sa nepokúšal toto testovať; nezdá sa mi to hodné toho rizika.

Ale ak je toto moja obrana, potom môj čin nemožno brániť ako dobrý skutok, nie? Pretože som práve vymenoval úžitok, aký to má pre mňa.

Nuž... kto povedal, že som ten čin *obhajoval* ako nesebecký dobrý skutok? Bol to *sebecký* dobrý skutok. Ak obnoví moju silu vôle, alebo ak ma drží v altruistickej nálade, potom to má aj nepriamy prospech pre druhých (aspoň si to myslím). Mohli by ste samozrejme odpovedať, že nedôverujete sebeckým činom, ktoré sú údajne prospešné pre druhých ako „postranný motív“; ale ja by som mohol rovnako ľahko odpovedať, že podľa toho istého princípu by ste sa mali skrátka pozrieť priamo na pôvodný dobrý skutok a nie na *jeho* údajný postranný motív.

Môže mi to takto prejsť? Myslím tým, môže mi naozaj prejsť, keď to nazvem „sebecký dobrý skutok“ a stále z toho budem čerpať obnovenie sily vôle, namiesto aby som som cítil vinu za to, že som sebecký? Zrejme áno. Som prekvapený, že to funguje takto, ale funguje to. Pokiaľ im zaklopem, aby som im povedal o otvorenom kufri, a pokiaľ dotyčný povie: „Ďakujem!“, môj mozog má pocit, že urobil svoj úžasný dobrý skutok pre tento deň.

U vás to môže byť inak, samozrejme. Problémom pri snahe vytvoriť umenie obnovy sily vôle je, že sa zdá, že pre rôznych ľudí fungujú rôzne veci. (Čo znamená: Skúšame rôzne veci na úrovni povrchného javu, a nerozumieme hlbším pravidlám, ktoré by predpovedali aj tieto rozdiely.)

Ale ak zistíte, že ste po tejto stránke ako ja – že sebecký dobrý skutok stále funguje – potom vám odporúčam, aby ste si *kupovali hrejivé pocity a utilony osobitne*. Nie naraz. Pokúsiť sa urobiť obidvoje zároveň znamená iba to, že ani jedno nedopadne dobre. Ak vám záleží aj na spoločenskom postavení, kupujte si aj spoločenské postavenie osobitne!

Keby som mal dať radu nejakému čerstvému miliardárovi, ktorý vstupuje do oblasti charity, moja rada by znela asi nejako takto:

- Aby ste si kúpili hrejivé pocity, nájdite nejakú ťažko pracujúcu ale chudobou sužovanú ženu, ktorá sa chystá zanechať štátnu vysokú školu, pretože jej manželovi skrátili úväzok, a osobne, ale anonymne, jej dajte šek na 10 000 dolárov. Opakujte, koľkokrát chcete.
- Aby ste si kúpili spoločenské postavenie medzi svojimi priateľmi, darujte 100 000 dolárov na aktuálne najatraktívnejšiu X-Prize, alebo hocijakú inú charitu, ktorá vyzerá, že ponúka najštýlovejšiu vec za najnižšiu cenu. Urobte okolo toto veľké haló, ukážte sa na ich tlačovej konferencii, a chváľte sa tým ďalších päť rokov.
- Potom – s absolútne chladnokrvným výpočtom – bez necitlivosti voči rozsahu alebo odporu k nejednoznačnosti – bez záujmu o spoločenské postavenie alebo hrejivé pocity – objavte nejakú

spoločnú schému na premenu výsledkov na utilony, a pokúste sa vyjadriť neistotu v percentách pravdepodobnosti – a nájdite charitu, ktorá ponúka najviac očakávaných utilonov za dolár. Darujte jej toľko peňazí, koľko ste chceli dať na charitu, dokiaľ ich hraničná efektivita neklesne pod efektivitu nasledujúcej charity na zozname.

Ďalej by som tomuto miliardárovi poradil, aby míňal na utilony prinajmenšom povedzme 20-krát toľko čo míňa na hrejivé pocity – 5 % navyše, aby ste sa udržali ako altruista vyzerá rozumne, a ja ako váš nezaujatý sudca by som nemal problém *odsúhlasiť* hrejivé pocity voči takémuto veľkému násobiteľovi. S tou výhradou, že ten pôvodný hrejivý čin by naozaj mal pomáhať a nie aktívne škodiť.

(Nakupovanie spoločenského *postavenia* mi pripadá, že s altruizmom v podstate nesúvisí. Ak vám darovanie peňazí na X-Prize získa viac obdivu od vašich kamarátov než nákup rovnako drahej lode, potom naozaj nemá zmysel kupovať tú loď. Akurát tieto peniaze zaúčtujte do stĺpca „robenie dojmu na kamarátov“ a uvedomujte si, že toto nie je stĺpec „altruizmus“.)

Ale hlavná lekcia je, že všetky tri tieto veci – hrejivé pocity, spoločenské postavenie, a očakávané utilony – si možno kupovať *omnoho* efektívnejšie, keď si ich kupujete *samostatne*, optimalizujete zakaždým na jednu vec. Napísať šek na 10 000 000 dolárov na charitu proti rakovine prsníka – hoci je to omnoho chvályhodnejšie než minúť tých istých 10 000 000 dolárov na, čojaviem, párty alebo niečo také – vám nedá tú koncentrovanú eufóriu z osobnej prítomnosti, keď zmeníte celý život jedného človeka, pravdepodobne sa to ani zďaleka nebude *podobať*. Nedá vám to takú tému na párty ako darovanie na niečo sexi ako je X-Prize – možno iba krátke prikývnutie od iných boháčov. A keby ste odhodili všetky starosti o hrejivé pocity a spoločenské postavenie, pravdepodobne existuje aspoň *tisíc* podfinancovaných charít, ktoré by za desať miliónov dolárov dokázali vyprodukovať *rádovo* viac utilonov. Pokúšať sa optimalizovať na všetky tri kritériá v jednom ťahu akurát zabezpečí, že žiadne z nich nebude naozaj dobre optimalizované – iba nejasné potlačenie vo všetkých troch rozmeroch.

Samozrejme, ak nie ste milionár alebo miliardár... potom nemôžete byť vo veciach rovnako *efektívni*, nemôžete tak ľahko nakupovať vo veľkom. Ale stále by som povedal – na hrejivé pocity si nájdite relatívne *lacnú* charitu s jasným, živým, ideálne osobným a priamym pomáhaním. Buďte dobrovoľníkom v kuchyni pre chudobných. Alebo získavajte hrejivé pocity tak, že držíte otvorené dvere malým starým paniam. Toto *potvrďte* svojimi inými snahami kupovať utilony, ale *nemýľte* si to s kupovaním utilonov. Spoločenské postavenie sa pravdepodobne najlacnejšie kupuje tak, že si kúpite pekné oblečenie.

A keď dôjde na nákup očakávaných utilonov – vtedy, samozrejme, držte hubu a počítajte.

*→

326. Ľahostajnosť diváka

Efekt diváka, známy aj ako ľahostajnosť diváka, je že väčšie skupiny v prípade naliehavej situácie reagujú s menšou pravdepodobnosťou – nie len jednotlivo, ale kolektívne. Dajte pokusnú osobu samu do miestnosti a začnite popod dvere púšťať dym. 75 % pokusných osôb odíde, aby to nahlásilo. Teraz dajte do miestnosti *tri* pokusné osoby – skutočné pokusné osoby, z ktorých žiadna nevie, o čo ide. Iba v 38 % prípadov vôbec *niekto* nahlási dym. Dajte pokusnú osobu s dvoma komplicmi, ktorí dym ignorujú, a nahlási ho iba v 10 % prípadov – dokonca zostane v miestnosti, kým sa celá nezadymí.[299]

Podľa štandardného modelu sú dva hlavné poháňače ľahostajnosti diváka sú:

- *Rozptýlenie zodpovednosti* – každý dúfa, že niekto iný sa ako prvý postaví a znesie prípadné náklady konania. Keď nikto nekoná, byť časťou davu poskytuje výhovorku a zmenšuje šancu, že vás niekto osobne bude brať na zodpovednosť za výsledky.
- *Pluralistická nevedomosť* – ľudia sa snažia *vyzerať* pokojne zatiaľ čo hľadajú náznaky, a vidia... že ostatní vyzerajú pokojne.

→ http://lesswrong.com/lw/6z/purchase_fuzzies_and_utilons_separately/

299 Bibb Latané and John M. Darley, „Bystander ‚Apathy,'“ [„Ľahostajnosť“ diváka] *American Scientist* 57, no. 2 (1969): 244–268, http://www.jstor.org/stable/27828530.

Cialdini:[300]

> Veľmi často naliehavá situácia nevyzerá ako zrejmá naliehavá situácia. Je muž ležiaci v parku obeťou infarktu alebo opilec vyspávajúci opicu? … V prípadoch takejto neistoty je prirodzený sklon obzrieť sa a hľadať náznaky v tom, čo robia druhí. Podľa toho, ako reagujú druhí svedkovia, sa môžeme dozvedieť, či daná udalosť je alebo nie je naliehavá. Je však ľahké zabudnúť na to, že všetci ostatní svedkovia udalosti pravdepodobne tiež hľadajú sociálne indície. Pretože sa všetci pred druhými radšej tvárime vyrovnane a pokojne, pravdepodobne budeme tieto indície hľadať pokojne, krátkymi maskovanými pohľadmi na ľudí okolo nás. Preto každý pravdepodobne uvidí, že všetci ostatní sa tvária flegmaticky a nekonajú.

Cialdini odporúča, že ak sa niekedy ocitnete v situácii, kde budete naliehavo potrebovať pomoc, ukáže na *jedného* diváka a požiadajte ho o pomoc – aby bolo veľmi jasné, na koho ukazujete. Pamätajte si, že *celá* skupina dokopy pomôže s menšou pravdepodobnosťou ako jeden jednotlivec.

Trochu som premýšľal nad evolučnou psychológiou efektu diváka. Predpokladajme, že v pravekom prostredí väčšina ľudí vo vašej tlupe boli pravdepodobne aspoň trochu vaši príbuzní – dosť na to, aby sa ich oplatilo zachrániť, keby ste boli jediný, kto to môže urobiť. Ale ak sú tam ešte dvaja ďalší, potom *prvý* konajúci človek znáša náklady, zatiaľ čo zvyšní dvaja zožnú *genetický* prospech z toho, že bol zachránený ich príbuzný. Mohla tam byť súťaž, kto vydrží najdlhšie čakať?

Pokiaľ som sledoval túto myšlienkovú líniu, nepripadá mi ako dobré vysvetlenie – v bode, kde už celá skupina nedokáže konať, by mal byť schopný prísť gén, ktorý pomáha ihneď, myslím si. Experimentálny výsledok neznie, že trvá dlho, kým príde pomoc, ale že pomoc vôbec nepríde: ak je geneticky výhodné pomôcť, keď ste jediným človekom, ktorý to môže urobiť (ako sa *deje* v týchto pokusoch), potom by skupinová rovnováha nemala spočívať v tom, že *nikto* nepomôže (ako sa deje v týchto pokusoch).

Nemyslím si teda, že súťaž v otáľaní je uveriteľným evolučným vysvetlením. Skôr si predstavujem, že sa pozeráme na problém, ktorý nie je praveký. Ak pokusné osoby domnelú obeť naozaj *poznajú*, šanca na pomoc prudko stúpa (čiže, nepozeráme sa na korelát pomoci skutočného člena našej tlupy). Ak si správne spomínam, ak sa pokusné osoby poznali *navzájom*, šanca na konanie tiež stúpla.

Môže tu hrať rolu aj nervóznosť pri jednaní na verejnosti. Ak má Robin Hanson pravdu ohľadom evolučnej úlohy „hrče v krku“, potom byť *prvým*, kto koná v naliehavej situácii, môže byť brané aj ako nebezpečný nárok na vysoké postavenie. (Keď sa nad tým zamyslím, v skutočnosti si nespomínam, že by niekto v analýze efektu diváka diskutoval o *hanblivosti*, ale možno je to len moja slabá pamäť.)

Možno efekt diváka vysvetliť v prvom rade rozptylom morálnej zodpovednosti? Mohli by sme byť cynickí a tvrdiť, že ľudia sa najviac zaujímajú o to, *aby ich nikto neobviňoval* za to, že nepomohli, než že by mali nejakú pozitívnu túžbu pomôcť – že si hlavne želajú vyhnúť sa antihrdinstvu a možnej odplate. Niečo také tu ľahko môže prispievať, ale dve pozorovania, ktoré to zmierňujú, sú (a) pokusné odoby nehlásili dym prenikajúci popod dvere, hoci mohol ľahko reprezentovať hrozbu pre nich osobne, a (b) povedať ľuďom o efekte diváka znižuje efekt diváka, hoci nie je väčšia pravdepodobnosť, že by za to boli verejne braní na zodpovednosť.

V skutočnosti je efekt diváka jedným z hlavných prípadov, na ktoré si narýchlo spomínam, kde povedať ľuďom o skreslení naozaj vyzerá, že ho prudko znižuje – možno preto, lebo vhodný spôsob kompenzovania je taký zrejmý, a nie je jednoduché to *prehnať* s kompenzáciou (ako keď sa snažíte napríklad prispôsobiť svoju kalibráciu). Mali by sme teda byť opatrní, aby sme neboli príliš cynickí ohľadom dôsledkov efektu diváka a rozptylu zodpovednosti, ak interpretujeme činy jednotlivcov ako chladný, vypočítavý pokus vyhnúť sa verejnej kritike. Zdá sa, že ľudia sa aspoň niekedy *sami* chopia zodpovednosti, keď si uvedomia, že sú jediní, kto vie dosť o efekte diváka, aby bolo pravdepodobné, že bude konať.

300 Cialdini, *Influence*. [Vplyv: Veda a prax]

Ale zaujímalo by ma, čo by sa stalo, keby ste vedeli, že ste súčasťou davu, kde *každému* povedali o efekte diváka...

*→

327. Kolektívna ľahostajnosť a internet

V predchádzajúce kapitole som opísal efekt diváka, alias ľahostajnosť diváka: v danej konštantnej problémovej situácii má *skupina* divákov v skutočnosti *menšiu* pravdepodobnosť, že zasiahne, než *osamelý* divák. Štandardné vysvetlenie tohto výsledku je pomocou pluralistickej nevedomosti (ak nie je jasné, či je situácia naliehavá, každý sa snaží *vyzerať* pokojne a vrhá pohľady na ostatných divákov, a vidí, že ostatní *vyzerajú* pokojne) a rozptylu zodpovednosti (každý dúfa, že niekto iný začne konať ako prvý; byť časťou davu zmenšuje tlak na jednotlivca až do bodu, kde nekoná nikto).

Čo môže byť príznakom toho, ako naše koordinačné mechanizmy lovcov-zberačov porazilo moderné prostredie. V pravekom prostredí ste zvyčajne nevytvárali pracovné skupiny s cudzincami; boli to zvyčajne ľudia, ktorých ste poznali. A naozaj, keď sa pokusné osoby navzájom poznajú, efekt diváka sa zmenší.

Ja viem, že toto je úžasné a revolučné pozorovanie, a dúfam že žiaden z mojich čitateľov teraz nezomrie následkom šoku, keď poviem toto: zdá sa, že ľudia majú problém reagovať konštruktívne na problémy, na ktoré natrafia cez internet.

Možno preto, lebo naše vrodené inštinkty na koordináciu nie sú naladené na to, že:

- Sme súčasťou skupiny cudzincov. (Keď sa všetky pokusné osoby navzájom poznajú, efekt diváka sa zmenší.)
- Sme súčasťou skupiny neznámej veľkosti, s cudzincami s neznámou totožnosťou.
- Nie sme navzájom vo fyzickom kontakte (alebo vizuálnom kontakte); nedokážeme si vymieňať významné pohľady.
- Nekomunikujeme v reálnom čase.
- Nie sme jeden druhému zaviazaní za iné formy pomoci; nie sme spolu závislí na našej skupine.
- Sme chránení pred poškodením povesti alebo strachom z poškodenia povesti svojou vlastnou zrejmou anonymitou; nepozerá sa na vás viditeľne nikto, voči ktorému by vaša povesť mohla utrpieť v dôsledku nekonania.
- Sme časťou veľkého kolektívu iných nekonajúcich; nikto nebude obviňovať vás osobitne.
- Nepočujeme vyslovenú prosbu o pomoc.

A tak ďalej. Nemám na tejto problém žiadne skvelé riešenie. Ale je to jedna z tých vecí, nad ktorými by som chcel, aby potenciálni zakladatelia dotcomov uvažovali explicitne, namiesto toho, aby riešili, ako pridať ovečky do Facebooku. (Áno, pozerám *na vás*, Hacker News!) Existujú webové aplikácie na online aktivizmus, ale majú sklon byť v duchu: *podpíšte túto petíciu! hurá, podpísali ste niečo!* namiesto: *Ako môžeme prekonať efekt diváka, obnoviť motiváciu, a pracovať s prirodzenými inštinktmi koordinácie skupiny cez internet?*

Pár vecí, ktoré mi napadli:

- Dajte na internet video, ako niekto žiada o pomoc.
- Dajte tam mená a fotky, alebo, ak sa dá, dokonca aj krátke videá ľudí, ktorí pomohli ako *prví* (alebo majte nejaký algoritmus priority v redditovskom štýle, závisiaci od kombinácie množstva pomoci a nedávnosti).
- Dajte pomáhajúcim video poďakovania od zakladateľov danej kauzy, ktoré si môžu umiestniť na svoju stránku „ľudia, ktorým som pomohol“, ktorá by s dostatočnou štandardizáciou mohla byť čiastočne alebo celá automaticky zostavená a ľahko vložiteľná do ich osobnej webovej stránky alebo do konta na Facebooku.

→ http://lesswrong.com/lw/9j/bystander_apathy/

- Nájdite *nie otravný* systém na „Povedz svojim priateľom o kauze X“; umožnite referentské kódy v linkách; potom ukážte ľuďom, koľkých ďalších oslovili (koľko ľudí, ktorí sa sem prvýkrát dostali pomocou referentského kódu X naozaj prispelo alebo urobilo niečo).
- (Všetko horeuvedené platí nielen pre dary, ale aj pre open-source projekty, do ktorých ľudia prispeli kódom. Alebo ak ľudia naozaj nechcú nič iné ako podpisy na petíciu, tak potom aj pre podpisy. Existujú aj iné spôsoby pomoci okrem peňazí – ale peniaze sú zvyčajne najefektívnejšie. Hlavná vec je, aby daná forma pomoci bola overiteľná cez internet.)
- Uľahčite ľuďom poskytovanie finančných odmien za podúlohy, kde sa výkon dá overiť.
- Ale hlavne som vám práve predložil otvorený, nevyriešený problém: umožnite / uľahčite skupinám cudzích ľudí spojiť sa do efektívnej pracovnej skupiny cez internet, napriek zvyčajným spôsobom zlyhania a štandardným dôvodom, prečo je toto nie-praveký problém. Pomyslite na tú starú štatistiku, že Wikipédia predstavuje 1/2000 času stráveného iba v USA sledovaním televízie. Tu vonku je celkom dosť paliva, len keby existovalo niečo také ako efektívny motor...

*→

328. Postupný pokrok a údolie

Rozumnosť je systematizované vyhrávanie.

„Ale,“ namietate, „rozumný človek *nie vždy* vyhrá!“

Čo tým myslíte? Myslíte tým, že každý druhý týždeň niekto, kto si kúpil žreb v lotérii so zápornou očakávanou hodnotou, vyhrá lotériu a stane sa bohatším než vy? To nie je *systematická* strata; to je selektívne spravodajstvo médií. Zo štatistického pohľadu výhercovia lotérie neexistujú – nebyť selektívneho spravodajstva, za celý svoj život by ste nestretli ani jedného.

Dokonca aj dokonale rozumní činitelia môžu prehrať. Akurát nemôžu *vopred vedieť*, že prehrajú. Nemôžu *očakávať, že sa im bude dariť horšie* než ľubovoľnej inej vykonateľnej stratégii, lebo inak by ju jednoducho vykonali.

„Nie,“ hovoríte, „hovorím o tom, ako zakladatelia firiem zbohatnú vďaka tomu, že veria v seba a svoje myšlienky silnejšie než by mohol ľubovoľný rozumný človek. Hovorím o tom, že veriaci ľudia sú šťastnejší...“

Aha. Nuž, tu je tá vec: *Inkrementálny* krok smerom k rozumnosti, ak je výsledok ešte stále po iných stránkach nerozumný, nemusí priniesť *inkrementálne* viac vyhrávania.

Vety o optimalite, ktoré máme v teórii pravdepodobnosti a teórii rozhodovania, sú pre *dokonalú* teóriu pravdepodobnosti a teóriu rozhodovania. Neexistuje žiadna podobná veta, ktorá by hovorila, že ak začnete z nejakého chybného počiatočného stavu, každá *inkrementálna* úprava algoritmu, ktorá posúva túto štruktúru bližšie k ideálu, musí priniesť *inkrementálne* zlepšenie výkonu. Toto nebolo dokázané, pretože to v skutočnosti nie je pravda.

„Takže,“ hovoríte, „aký má potom zmysel snažiť sa byť rozumnejší? Ten dokonalý ideál nedosiahneme. Nemáme teda žiadnu záruku, že naše kroky vpred pomôžu.“

Nemáte ani žiadnu záruku, že vám krok *vzad* pomôže vyhrať. Záruky v tomto svete neexistujú; ale na rozdiel od bežných mýtov, usudzovanie v podmienkach *neistoty* je to, o čom je celá rozumnosť.

„Ale my tu máme niekoľko prípadov, založených buď na hmlisto rozumne znejúcom uvažovaní, alebo na údajoch z ankety, kde to vyzerá, že inkrementálny krok v rozumnosti nám uškodí. Ak je to naozaj celé o vyhrávaní – ak musíš chrániť niečo dôležitejšie než nejaký rituál poznávania – *prečo* potom urobiť tento krok?“

Aha, tak *teraz* sme sa dostali k podstate.

Nevyhnutne nemôžem odpovedať za každého, ale...

Mojím prvým dôvodom je, že v profesionálnom živote mám do činenia s hlboko zmätenými problémami, ktoré sú veľmi náročné na presnosť myslenia. Jedna malá chyba vás môže zviesť z cesty na

→ http://lesswrong.com/lw/9m/collective_apathy_and_the_internet/

celé roky, a za rohom čakajú ešte horšie tresty. Nezlepšená úroveň výkonu *nestačí*; mám na výber buď robiť veci lepšie, alebo sa vzdať a ísť domov.

„Ale to si len *ty*. Nie všetci vedieme takýto život. Čo ak skúšaš iba nejakú bežnú ľudskú úlohu, ako je založenie internetovej firmy?“

Môj druhý dôvod je, že sa pokúšam potlačiť nejaké stránky svojho umenia ďalej, než som videl urobené. *Neviem*, kam tieto vylepšenia vedú. To, čo stratíte, keď neurobíte krok vpred, nie je iba tento *jeden krok*, ale sú to všetky *ďalšie* kroky vpred, ktoré ste ďalej mohli urobiť. Robin Hanson má porekadlo: Problémom pri padaní zo schodov nie je pád z výšky prvého schodu, ale to, že pád o jeden schod vedie k padaniu o ďalší schod. Podobne, odmietnutím vyliezť o jeden krok vyššie strácate nielen výšku tohto kroku, ale výšku celého schodiska.

„Ale opäť – to si len ty. Nie všetci sa pokúšame potlačiť umenie do nezmapovaného územia.“

Môj tretí dôvod je, že keď si raz uvedomím, že som bol oklamaný, nedokážem len tak zavrieť oči a predstierať, že som to nevidel. Tento krok dopredu *som už urobil*; prečo to popierať sám pred sebou? Nedokázal by som veriť v Boha, ani keby som to skúšal, rovnako ako by som nedokázal veriť, že obloha nado mnou je zelená, zatiaľ čo by som sa pozeral priamo na ňu. Ak *viete* všetko, čo potrebujete vedieť, aby ste vedeli, že ste na tom lepšie, ak sa klamete, už je príliš neskoro sa klamať.

„Ale toto uvedomenie je *nezvyčajné*; väčšina ľudí to má s doublethinkom ľahšie, pretože si neuvedomujú, že sa to nedá. *Ty* tu chodíš a snažíš sa aktívne podporovať kolaps doublethinku. *Ty*, z vyššieho pozorovacieho miesta, môžeš vedieť dosť na to, aby očakával, že ich to urobí menej šťastnými. Je to teda sadistická túžba ublížiť tvojim čitateľom, alebo čo?“

Potom nakoniec odpoviem, že moja doterajšia skúsenosť – dokonca aj v tejto oblasti púhych ľudských možností – *naznačuje*, že keď si raz v sebe trochu upracete a nerobíte *až toľko* veľa iných vecí zle, snaha o rozumnosť v skutočnosti vašu situáciu *zlepší*. Tá dlhá cesta vedie von z údolia a vyššie než predtým, dokonca už v ľudských krajoch.

Čím viac viem o nejakej konkrétnej stránke Umenia, tým viac môžem vidieť, že je to tak. Ako som predtým poznamenal, moje eseje možno neodrážajú to, aké by mohlo byť skutočné bojové umenie rozumnosti, pretože som sa sústredil iba na odpovedanie na zmätené otázky – nie na boj s akráziou, koordinovanie skupín, alebo šťastie. V oblasti odpovedania na zmätené otázky – oblasti, kde som najintenzívnejšie cvičil Umenie – to teraz vyzerá *drvivo* zrejmé, že ktokoľvek by si myslel, že na tom bude lepšie, ak „zostane optimista ohľadom riešenia problému“, by bol zadupaný *do zeme*. *Bežným študentom.*

Keď ide o to udržať sa motivovaný alebo byť šťastný, nemôžem zaručiť, že niekto, kto stratí svoje ilúzie, bude na tom lepšie – pretože moje vedomosti o týchto stránkach rozumnosti sú stále hrubé. Ak tieto časti Umenia boli vyvíjané systematicky, ja o tom neviem. Ale dokonca aj tu som podstúpil istú značnú námahu, aby som rozptýlil polo-rozumné polo-pomýlené predstavy, ktoré by sa mohli začiatočníkovi dostať do cesty, ako je predstava, že rozumnosť odporuje cíteniu, alebo predstava, že rozumnosť odporuje hodnote, alebo predstava, že sofistikovaní myslitelia by mali byť úzkostní a cynickí.

A ak, ako dúfam, niekto pôjde ďalej vo vývoji umenia boja s akráziou alebo dosahovania mentálneho zdravia rovnako dôkladne, ako som ja vyvinul umenie odpovedania na nemožné otázky, plne očakávam, že tí, ktorí sa zabalia do svojich ilúzií, nebudú *schopní* súťažiť. Medzičasom – iným sa môže dariť lepšie než mne, ak je šťastie ich najdrahšia túžba, lebo ja som tu investoval málo úsilia.

Pripadá mi ťažké uveriť, že *optimálne* motivovaný jednotlivec, *najsilnejší* podnikateľ, akým sa človek môže stať, je stále zabalený v deke upokojujúcej prehnanej sebadôvery. Myslím si, že pravdepodobne túto deku vyhodil von oknom a zorganizoval si svoju myseľ o trochu *inak*. Pripadá mi ťažké veriť, že to najväčšie šťastie, aké len môžeme zažiť, dokonca aj v oblasti ľudských možností, zahŕňa malé vedomie číhajúce v kútiku vašej mysle, že to celé je lož. Radšej by som stavil svoje nádeje na neurofeedback alebo zenovú meditáciu, hoci som ani jedno z toho neskúšal.

Ale nemožno popierať, že toto je veľmi reálna téma z veľmi reálneho života. Vezmite si tieto dva komentáre z Less Wrong:

Budem úprimný – môj život sa otočil prudko nadol, odkedy som dekonvertoval. Moja veriaca priateľka, ktorú som veľmi miloval, nemohla zniesť túto moju zmenu, a po šiestich mesiacoch bolestivého kolísania, ma opustila kvôli kolegovi z práce. Odvtedy prešlo ďalších šesť mesiacov, a stále mám zlomené srdce, cítim sa mizerne, neviem sa sústrediť, a som *extrémne* neefektívny.

Možno je toto príkladom toho údolia zlej rozumnosti, o ktorom hovorí PhilGoetz, ale ja aj tak hodnotím svoju terajšiu situáciu vyššie na mojom rebríčku preferencií než šťastie s falošnými názormi.

A:

Máš môj súcit: toto sa mi stalo asi pred 6 rokmi (hoci našťastie bez toľkého viditeľného kolísania).

Moja sestra, ktorá mala nejaký tréning v Kognitívne behaviorálnej terapii, mi pripomínala, že vzťahy vznikajú a zanikajú každú chvíľu, a keďže nie som neatraktívny a neutiahol som sa do kláštornej samoty, nie je rozumné myslieť si, že budem sám po celý zvyšok svojho života (ukázalo sa, že má pravdu). Toto mi pomáhalo v časoch, keď ma moje pocity celkom zmáhali.

Takže – v praxi, v skutočnom živote, ako triezvy fakt – tieto prvé kroky naozaj môžu byť bolestivé. A potom sa veci naozaj môžu zlepšiť. A naozaj neexistuje žiadna *záruka*, že na konci budeš na tom lepšie než predtým. Aj keď v princípe táto cesta musí ísť ďalej, nie je žiadna záruka, že nejaký konkrétny človek zájde tak ďaleko.

Ak nedávaš *prednosť* pravde pred šťastím s falošnými názormi...

No... *a* ak nerobíš nič mimoriadne neisté alebo mätúce... a ak si nekupuješ žreby z lotérie... a ak si už prihlásený na kryoniku, čo je veľmi mätúci test ohňom rozumnosti s nečakanými, mimoriadne vysokými stávkami, ktorý ilustruje vlastnosti „čiernych labutí“, keď sa snažíme stávkovať o nevedomosti *v* nevedomosti...

Potom nemáš *záruku*, že podniknúť všetky inkrementálne kroky k rozumnosti, ktoré dokážeš nájsť, ťa zanechá v lepšej situácii. Ale tie hmlisto dôveryhodne znejúce argumenty proti strácaniu svojich ilúzií vo všeobecnosti zvažujú *iba* jeden jediný krok, bez predpokladu ďalších krokov, bez odporúčania nejakého pokusu, ako získať späť všetko, čo bolo stratené, a ísť ešte ďalej. Dokonca aj tie prieskumy porovnávajú priemerného veriaceho a priemerného ateistu, nie špičkových teológov a špičkových racionalistov.

Ale ak ti nezáleží na pravde – *a* nemáš čo chrániť – *a* nepriťahuje ťa predstava posúvania svojho umenia tak ďaleko, ako sa len dá – *a* ak sa zdá, že tvoj terajší život sa vyvíja dobre – *a* ak cítiš, že tvoja duševná vyrovnanosť závisí od ilúzií, na ktoré by si radšej nemyslel...

Potom pravdepodobne nečítaš tento text. Ale ak áno, potom asi... no... (a) prihlás sa na kryoniku, a potom (b) *prestaň čítať Less Wrong skôr než sa tvoje ilúzie zrútia! UTEKAJ PREČ!*

* →

329. Bayesovci verzus barbari

Predtým:

Povedzme, že máme dve skupiny vojakov. V skupine 1, pešiaci nevedia o taktike a stratégii; iba seržanti vedia o taktike, a iba dôstojníci vedia o stratégii. V skupine 2 každý na každej úrovni vie všetko o taktike a stratégii.

→ http://lesswrong.com/lw/7k/incremental_progress_and_the_valley/

> Mali by sme očakávať, že skupina 1 porazí skupinu 2, pretože skupina 1 sa bude riadiť príkazmi, zatiaľ čo každý v skupine 2 príde *s lepším nápadom* než bol hocijaký rozkaz, ktorý dostal?
>
> V tomto prípade by som musel pochybovať o tom, nakoľko skupina 2 naozaj rozumie vojenskej teórii, pretože je *základnou* poučkou, že nekoordinovaný dav bude zmasakrovaný.

Predstavte si, že krajinu racionalistov napadne krajina Zlých Barbarov, ktorí nevedia nič o teórii pravdepodobnosti a teórii rozhodovania.

Existuje určitý názor na „rozumnosť“ alebo „racionalizmus“, ktorý by povedal niečo takéto:

„Samozrejme, racionalisti prehrajú. Barbari veria v posmrtný život, kde budú odmenení za odvahu; takže sa vrhnú do boja bez váhania a ľútosti. Vďaka svojim afektívnym špirálam smrti ohľadom svojej Kauzy a Veľkého Vodcu Boba budú ich bojovníci poslúchať rozkazy, a ich obyčania doma budú nadšene a naplno produkovať produkovať pre vojnu; každý, koho prichytia, že rozkráda alebo sa ulieva, bude upálený na hranici v súlade s barbarskou tradíciou. Budú veriť vo svoje dobro a nenávidieť nepriateľa silnejšie, než by mohol akýkoľvek príčetný človek, čím sa spoja do pevnej skupiny. Medzičasom si racionalisti uvedomia, že neexistuje žiadna predstaviteľná odmena pre tých, čo padnú v boji; budú si želať, aby bojovali druhí, ale nebudú chcieť bojovať sami. Dokonca aj keď nájdete vojakov, ich civili nebudú natoľko spolupracovať: Dokiaľ nejaká *jedna* klobáska nepovedie s takmer istotou k zrúteniu vojnového úsilia, budú si chcieť túto klobásku nechať pre seba, a neposkytnú toľko, koľko by mohli. Bez ohľadu na to, aká uhladená, elegantná, civilizovaná, produktívna a nenásilná bola ich kultúra na začiatku, nedokážu odolať invázii Barbarov; príčetná diskusia nie je protiváhou šialencovi s penou v ústach a puškou v ruke. Nakoniec Barbari vyhrajú preto, lebo *chcú* bojovať, *chcú* racionalistom ublížiť, *chcú* ich dobyť a celá ich spoločnosť je zjednotená pre dobývanie; záleží im na tom viac než by mohlo hocijakému príčetnému človeku.“

Vojna nie je zábava. Ako zistilo mnoho ľudí od úsvitu zaznamenávaných dejín, ako zistilo mnoho ľudí pred úsvitom zaznamenávaných dejín, a ako nejaká spoločnosť zisťuje práve teraz v nejakej smutnej malej krajine, ktorej vnútorné utrpenie sa už ani nedostáva na titulné stránky novín.

Vojna nie je zábava. *Prehrať* vojnu je ešte menej zábavné. A už od dávnych čias sa hovorilo: „Ak chceš mať mier, pripravuj sa na vojnu.“ Tvoji súperi nemusia veriť, že *vyhráš*, že ich dobyješ; ale musia veriť, že budeš bojovať dosť na to, aby im to nestálo za to.

Môžete teda vidieť, že keby bolo naozaj osudom „racionalistov“ prehrať v každej vojne, nemohol by som potom s čistým svedomím odporúčať, aby si široká verejnosť osvojila „rozumnosť“.

Toto je pravdepodobne jedna z najškaredších tém, na akú mám v úmysle diskutovať na LW. Vojna nie je čistá. Súčasné vysoko technické armády – čím myslím armádu USA – sú jedinečné v tom, akou neporovnateľne väčšou silu dokážu udrieť na protivníkov, čo umožňuje z historického hľadiska výnimočný stupeň obavy o nepriateľove straty na životoch, a na životoch civilov.

Vyhrať vo vojne neznamenalo vždy odhodiť *všetku* morálku. Vojny boli vybojované aj bez použitia mučenia. Z nezábavnosti vojny nevyplýva, napríklad, že klásť prezidentovi otázky je nevlastenecké. Sme zvyknutí, že sa „vojna“ zneužíva ako zámienka pre zlé správanie, pretože v nedávnej histórii USA to *je* prakticky presne to, na čo sa používala...

Ale obrátená hlúposť nie je inteligencia. A obrátené zlo tiež nie je inteligencia. Zostáva pravdou, že *skutočnú* vojnu nemožno vybojovať uhladenou slušnosťou. Ak sa „racionalisti“ nedokážu pripraviť na ten myšlienkový šok, potom Barbari naozaj vyhrajú; a títo „racionalisti“... nechcem povedať, že „si zaslúžili prehrať“. Ale zlyhajú v tomto teste existencie ich spoločnosti.

Dovoľte mi začať odhodením predstavy, že ideálni rozumní činitelia *v princípe* nemôžu bojovať vo vojne, pretože každý z nich by bol radšej civilom než vojakom.

Ako som už dosť dlho vysvetľoval, ja si v Newcombovom probléme vyberám jednu krabicu.

Konzistentne s tým, *neverím*, že ak voľby skončili pomerom hlasov 100 000 k 99 998, potom všetci títo voliči boli nerozumní, keď vynakladali úsilie ísť do volebnej miestnosti, pretože „keby som ja zostal

doma, výsledok by to nezmenilo“. (Ani neverím, že ak voľby skončia 100 000 k 99 999, potom z týchto 100 000 ľudí boli *všetci*, jednotlivo, *osobne zodpovední* za výsledok.)

Konzistentne si aj myslím, že dve rozumné UI (ktoré používajú môj druh teórie rozhodovania), dokonca aj keby mali celkom rôzne funkcie úžitku a boli navrhnuté rôznymi tvorcami, budú spolupracovať v skutočnej Väzenskej dileme, ak majú spoločné poznanie navzájom svojho zdrojového kódu. (Alebo dokonca len spoločné poznanie vzájomnej *rozumnosti* v tom primeranom zmysle slova.)

Konzistentne, verím, že rozumní činitelia sú schopní koordinovať sa v skupinových projektoch vždy, keď je (očakávaný pravdepodobnostný) výsledok lepší než by bol bez takejto koordinácie. Spoločnosť činiteľov, ktorí používajú môj druh teórie rozhodovania, a majú spoločnú vedomosť o tomto fakte, skončí v paretovskom optime namiesto nashovskej rovnováhy. Ak sa všetci rozumní činitelia zhodnú na to, že je lepšie bojovať než vzdať sa, potom budú bojovať proti Barbarom namiesto toho, aby sa vzdali.

Predstavte si spoločenstvo sebamodifikujúcich UI, ktoré by kolektívne radšej bojovali než sa vzdali, ale jednotlivo by radšej boli civilmi než vojami. Jedným riešením je urobiť lotériu, nepredvídateľnú pre každého činiteľa, na výber vojakov. *Pred* spustením lotérie si všetky UI vopred zmenia svoj kód tak, že ak budú vylosované, budú bojovať tým najefektívnejsím spôsobom pre spoločenstvo – aj keby to znamenalo pokojne napochodovať v ústrety svojej smrti.

(Reflexívne konzistentná teória rozhodovania funguje rovnako, iba bez tej sebamodifikácie.)

Odpoviete: „Lenže v skutočnom svete ľudí, činitelia nie sú dokonael rozumní, ani nemajú spoločné poznanie navzájom o svojom zdrojovom kóde. Spolupráca vo Väzenskej dileme si vyžaduje podľa tvojej teórie rozhodovania isté podmienky (ktoré sa nezmestia do tejto poznámky) a tieto podmienky v skutočnom živote nie sú splnené.“

Odpovedám: Čistá, skutočná Väzenská dilema je v skutočnom živote neuveriteľne zriedkavá. V skutočnom živote zvyčajne máte ďalšie účinky – čo robíte, ovplyvňuje vašu povesť. V skutočnom živote väčšine ľudí do istej miery záleží na tom, čo sa stane druhým. A v skutočnom živote máte príležitosť vopred pripraviť motivačné mechanizmy.

A v skutočnom živote si *myslím*, že spoločenstvo ľudských racionalistov by dokázalo vyrobiť vojakov ochotných zomrieť na obranu svojho spoločenstva. Pokiaľ by sa deťom nehovorilo v školách, že ideálni racionalisti sa údajne majú navzájom podraziť vo Väzenskej dileme. Keby sa všeobecne myslelo – ako si to myslím ja, z celkom rovnakých dôvodov, prečo si v Newcombovom probléme vyberám jednu krabicu – že ak sa ľudia ako jednotlivci rozhodnú nebyť vojakmi, alebo ak sa vojaci rozhodnú zutekať, potom je to to isté ako rozhodnúť sa, aby Barbari vyhrali. Podľa tej istej teórie, v ktorej ak sa voľby vyhrajú pomerom 100 000 hlasov voči 99 998 hlasom, nemá zmysel, aby si každý voliť povedal „môj hlas nespôsobil žiaden rozdiel“. Keby sa hovorilo (lebo je to pravda), že funkcie úžitku nemusia byť solipsistické, a že rozumný činiteľ môže bojovať na smrť, pokiaľ mu dosť záleží na tom, čo chráni. Keby sa nehovorilo, že racionalisti majú za úlohu rozumne prehrať.

Keby toto bola kultúra a zvyky v racionalistickej spoločnosti, potom si myslím, že by sa *bežní ľudia* v tejto spoločnosti dobrovoľne prihlásili za vojakov. Napokon, zdá sa, že aj toto je do ľudí zabudované. Potrebujete iba zabezpečiť, že sa *do cesty nepostaví* kultúrny výcvik.

A ak sa mýlim, a ak nebudete mať dosť dobrovoľníkov?

Potom, dokiaľ ako celok stále dávajú prednosť boju pred vzdaním sa, majú príležitosť nastaviť motivačné mechanizmy, a vyhnúť sa skutočnej Väzenskej dileme.

Môžete mať lotérie, kto bude vybraný za bojovníka. Niečo ako v tom príklade, kde si UI menili svoj kód. Akurát, že ak „buď reflexívne konzistentný; rob to, čo si sa dopredu zaviazal robiť“ nie je dostatočná motivácia pre tých ľudí, ktorí vyhrajú v lotérii, potom...

...nuž, ešte pred skutočným prebehnutím lotérie by sme sa možno mohli všetci dohodnúť, že je dobrý nápad dať vylosovaným drogy, ktoré v nich vyvolajú extra odvahu, a zastreliť ich, ak budú utekať. Zohľadňujúc aj to, že my sami môžeme byť vylosovaní. Pretože *pred* prebehnutím lotérie je toto všeobecné pravidlo, ktoré nám dáva najvyššiu *očakávanú* šancu prežiť.

...ako som povedal: Skutočná vojna = nie zábava, prehrať vojnu = ešte menej zábavy.

Len aby bolo jasné, mimochodom, že nepodporujem povolávanie do zbrane tak, ako sa robí dnes. Tieto povolávania nie sú kolektívne pokusy obyvateľstva o pohyb z nashovskej rovnováhy do paretovského optima. Povolávanie je nástrojom kráľov, ktorí hrajú hry a potrebujú vojačikov. Trvám na to, že vojaci povolaní do Vietnamu, ktorí ušli do Kanady, boli v práve. Ale spoločnosť, ktorá sa považuje za príliš bystrú na to, aby mala kráľov, *nemusí* byť príliš bystrá na to, aby prežila. Dokonca aj keď útočia hordy Barbarov, a keď Barbari povolávajú do armády.

Bude rozumný vojak poslúchať príkazy? Čo ak veliaci dôstojník urobí chybu?

Vojaci pochodujú. Všetky nohy dopadajú na zem v rovnakom rytme. Možno aj proti ich vlastnému sklonu, pretože keby boli ľudia ponechaní na seba, kráčal by každý iným tempom. Lasery zostavené z ľudí. To je pochodovanie.

Ak je možné vyvinúť nejakú metódu skupinového rozhodovania, ktorá je *lepšia* ako keď kapitán dáva rozkazy, potom by rota rozumných vojakov mohla implementovať tento postup. Ak neexistuje žiadna overená metóda lepšia než kapitán, potom sa rota rozumných vojakov zaviaže poslúchať kapitána, dokonca aj proti svojim vlastným sklonom. A ak ľudia nie sú takí rozumní... potom ešte pred lotériou, všeobecné pravidlo, ktoré vám dáva najvyššiu osobnú pravdepodobnosť prežitia je strieľať vojakov, ktorí neposlúchnu rozkazy. Čím nechcem povedať, že tí, ktorí fragovali svojich vlastných dôstojníkov vo Vietname, neboli v práve; mohli totiž konzistentne dávať prednosť tomu, aby sa *nikto* nezúčastnil v povolávacej lotérii.

Lenže nekoordinovaný dav bude zmasakrovaný, a tak vojaci potrebujú *nejaký* spôsob ako robiť všetci to isté naraz pri dosahovaní rovnakého cieľa, napriek tomu, že keby boli ponechaní sami na seba, možno by pochodovali na všetky strany. Príkazy nemusia prichádzať od kapitána ako od veľkého kmeňového náčelníka, ale jednotné príkazy musia *odniekadiaľ* prísť. Spoločnosť, ktorej vojaci sú príliš chytrí na to, aby poslúchali príkazy, je spoločnosť, ktorá je príliš chytrá na to, aby prežila. Rovnako ako spoločnosť, ktorej ľudia sú príliš chytrí na to, aby sa *stali* vojakmi. Preto hovorím „chytrí“, čo zvyčajne používam ako kritiku, a nie „rozumní“.

(Ale myslím si, že je dôležitá otázka, či dokážeme vymyslieť metódu koordinovania malej skupiny, ktorá v praxi naozaj funguje lepšie než mať vodcu. Čím viac môžu ľudia dôverovať metóde skupinového rozhodovania – čím viac môžu veriť, že je to naozaj lepšie než keď si každý robí svoje – tým zladenejšie sa môžu správať dokonca aj v neprítomnosti vynútiteľných trestov za neposlušnosť.)

Hovorím toto všetko, hoci iste neočakávam, že racionalisti v dohľadnej dobe ovládnu nejakú krajinu, pretože si myslím, že to, čo veríme o spoločnosti „ľudí ako sme my“ má istý dopad na to, čo si myslíme sami o sebe. Ak veríte, že spoločnosť ľudí ako ste vy by bola príliš rozumná na to, aby dokázala dlhodobo prežiť... to je jeden druh sebaobrazu. A je celkom iný druh sebaobrazu, ak si myslíte, že spoločnosť ľudí celkom ako vy by mohla bojovať proti Zlým Barbarom a *vyhrať* – nie iba vďaka lepšej technológii, ale pretože by vašim ľuďom záležalo na sebe navzájom a na svojej kolektívnej spoločnosti – a pretože dokážu čeliť skutočnostiam vojny a nestratiť sa – a pretože by si spočítali, čo je skupinovo rozumné urobiť a zabezpečili by, že sa to stane – a pretože v pravidlách teórie pravdepodobnosti ani teórie rozhodovania nie je nič, čo by hovorilo, že sa nemôžete pre nejakú kauzu obetovať – a pretože ak naozaj *ste* bystrejší než Nepriateľ, a nie iba že si sami lichotíte, potom by ste mali byť schopní využívať slepé miesta, o ktorých si Nepriateľ nedovolí rozmýšľať – a pretože bez ohľadu na to ako veľmo sa nepriateľ vyhecuje pred bitkou, vy si myslíte, že možno koherentná myseľ, vnútorne nerozdelená, možno cvičiaca niečo podobné meditácii alebo samohypnóze, dokáže bojovať v praxi rovnako tvrdo ako niekto, kto teoreticky verí, že naňho čaká sedemdesiat dva panien.

Potom budete očakávať viac od seba *a od ľudí, ako ste vy, fungujúcich v skupinách*; a potom môžete vidieť sami seba ako niečo iné než kultúrnu slepú uličku.

Pozrite sa na to teda takto: Jeffreyssai by sa pravdepodobne Zlým Barbarom nevzdal, keby bojoval *sám*. Celá *armáda* majstrov *beisutsukai* by mala byť silou, s ktorou si *nikto* nezačína. To je motivačná vízia. Otázka je, ako presne to funguje.

*→

330. Vyvarujte sa optimalizácie druhých

Všimol som si vážny problém, že ašpirujúci racionalisti výrazne preceňujú svoju schopnosť optimalizovať životy druhých ľudí. A myslím si, že mám istú predstavu, ako tento problém vzniká.

Prečítate si devätnásť rôznych webových stránok radiacich ohľadom osobného zlepšenia – produktivita, diéta, šetrenie peňazí. A všetci títo autori znejú presvedčene a nadšeje ohľadom Ich Metódy, hovoria príbehy o tom, ako im to fungovalo a prinieslo *úžasné* výsledky...

Ale väčšina týchto rád znie tak falošne, že sa zdá, že ani nestoja za úvahu. Takže si vzdychnete, smutne dumajúc nad divokým detinským nadšením, ktoré si ľudia napohľad dokážu vyvinúť pre prakticky hocičo, bez ohľadu na to, aké hlúpe. Rady číslo 14 a 15 znejú zaujímavo, vyskúšate ich, ale... akosi... nie celkom... skrátka, hrozným spôsobom to zlyhá. Tá rada bola zlá, alebo vy ste to nedokázali, a tak či onak na tom nie ste o nič lepšie.

A potom si prečítate dvadsiatu radu – alebo dokonca objavíte dvadsiatu metódu, ktorá nebola na žiadnej z týchto stránok – a HVIEZDY NADO MNOU ONO TO TENTOKRÁT NAOZAJ FUNGUJE.

Konečne, konečne ste objavili tú *skutočnú* cestu, tú *správnu* cestu, tú naozaj *fungujúcu* cestu. A keď sa niekto iný dostane do toho druhu problémov, aké ste kedysi mali vy – nuž, tentokrát *viete*, ako mu pomôcť. Môžete ho ušetriť všetkého toho trápenia s čítaním si devätnástich zbytočných rád a preskočiť konečne k správnej odpovedi. Ako ašpirujúci racionalista ste sa už naučili, že väčšina ľudí nepočúva a zvyčajne sa ani nenamáhate – ale tento človek je váš kamarát, niekto, koho poznáte, niekto, komu dôverujete, že si vás vypočuje.

A tak mu položíte kamarátsky ruku na plece, pozriete mu priamo do očí, a poviete mu, ako na to.

Ja osobne dostávam tohto veľa. Pretože, viete... keď ste objavili postup, ktorý *naozaj funguje*... no, máte dosť rozumu na to, aby ste nebežali a nehovorili to svojím kamarátom a rodine. Ale musíte to skúsiť povedať Eliezerovi Yudkowskemu. On to *potrebuje*, a je celkom dobrá šanca, že *on* to pochopí.

Naozaj mi chvíľu trvalo, kým som to pochopil. Jedna z kritických udalostí bola, keď mi niekto z dozornej rady Výskumného ústavu strojovej inteligencie povedal, že nepotrebujem zvýšiť plat, aby som dorovnal infláciu – pretože môžem míňať omnoho menej peňazí na jedlo, ak použijem online službu so zľavami. Veril som tomu, pretože to bol kamarát, ktorému som dôveroval, a povedal mi to takým sebaistým tónom. Takže moja priateľka začala skúšať používať túto službu, a o pár týždňov neskôr to vzdala.

Teraz mi ide o toto: keby som nahrafil na presne rovnakú radu o používaní zliav na nejakom inom blogu, pravdepodobne by som jej ani nevenoval veľa pozornosti, len by som si ju prečítal a šiel ďalej. Dokonca aj keby to napísal Scott Aaronson alebo niekto podobný, o kom viem, že je inteligentný, stále by som si to prečítal a šiel ďalej. Ale keďže mi to osobne doručil kamarát, ktorého som poznal, môj mozog to spracoval inak – akoby mi bolo prezradené *tajomstvo*; a veru, takým tónom mi to aj bolo povedané. A až oneskorene som si uvedomil, že mi jednoducho bolo povedané, ako osobná rada, niečo, čo by za iných okolností bol iba článok na nejakom blogu; nemá to o nič väčšiu ani o nič menšiu pravdepodobnosť, že to bude pre mňa fungovať, než blog o produktivite, ktorý by napísal ľubovoľný iný inteligentný človek.

A pretože som sa stretol s veľmi mnohými ľuďmi, ktorí sa ma snažia optimalizovať, môžem dosvedčiť, že rady, ktoré dostávam, sú rovnako rozmanité ako blogosféra produktivity. Lenže ostatní nevidia túto rozmanitosť rád o produktivite ako náznak, že ľudia sa *líšia* v tom, ktorá rada pre nich funguje. Namiesto toho to vidia ako hromadu samozrejme zlých úbohých rád. A potom nakoniec objavia tú správnu cestu – tú, ktorá funguje, na rozdiel od všetkých tých zvyšných článkov na blogoch, ktoré nefungujú – a potom sa celkom často rozhodnú použiť to na optimalizáciu Eliezera Yudkowskeho.

→ http://lesswrong.com/lw/5f/bayesians_vs_barbarians/

Neberte to v zlom. Niekedy sú tie rady užitočné. Niekedy fungujú. „Zaseknutý uprostred s Bruceom“ - toto ma hlboko oslovilo. Možno sa ukáže, že je to tá najužitočnejšia vec, akú som zatiaľ čítal na *Less Wrong*, hoci to sa ešte len ukáže.

Ide mi len o to, že vaše úprimné osobné rady, tie úžasné veci, o ktorých ste zistili, že naozaj zázračne fungujú, nemajú o nič väčšiu a o nič menšiu pravdepodobnosť, že budú fungovať pre mňa, než náhodný osobný blog o sebazdokonaľovaní napísaný inteligentným autorom bude fungovať pre vás.

„Rôzne veci fungujú pre rôznych ľudí.“ Táto veta veta vám môže dať divný pocit; viem, že mne ho dáva. Pretože táto veta je nástroj, ktorý používa Epistemológia Temnej Strany, aby sa zaštítila pred kritikou, používaná podobne ako „Rôzne veci sú pravdivé pre rôznych ľudí“ (čo je jednoducho nepravda).

Ale dokiaľ nepochopíte zákony, ktoré sú takmer všeobecnými zovšeobecneniami, niekedy skončíte tak, že sa pohrávate s povrchnými trikmi, ktoré pre jedného človeka fungujú a pre druhého nie, bez porozumenia prečo, pretože nepoznáte všeobecné zákony, ktoré by diktovali, čo bude pre koho fungovať. A to najlepšie, čo môže urobiť, je pamätať na to a akceptovať „Nie“ ako odpoveď.

A *najmä* by ste mali dokázať akceptovať „Nie“ ako odpoveď, ak máte nad tým druhým *moc*. Moc je vo všeobecnosti veľmi nebezpečná vec, ktorá sa ohromne ľahko zneužíva, ani si nemusíte uvedomiť, že ju zneužívate. Existujú veci, ktoré môžete urobiť, aby ste si zabránili zneužívať moc, ale musíte ich naozaj urobiť, inak nebudú fungovať. Na *Overcoming Bias* je článok o tom, ako sa ukázalo, že byť v mocenskej pozícii znižuje vašu schopnosť vciťovať sa do druhých a rozumieť im, ale teraz ho neviem nájsť. Videl som racionalistu, ktorý si nemyslel, že má moc, a tak si nemyslel, že si musí dávať pozor, a bol prekvapený keď zistil, že sa ho druhí boja...

Je to ešte horšie, keď tento objav, ktorý pre nich funguje, vyžaduje trochu *sily vôle*. Potom, ak poviete, že pre vás to nefunguje, odpoveď je jasná a samozrejmá: ste skrátka *lenivý* a musia na vás vyvinúť nejaký *tlak*, aby vás primäli urobiť tú *správnu* vec, tú radu, o ktorej zistili, že naozaj funguje.

Niekedy – predpokladám – sú ľudia naozaj leniví. Ale buďte veľmi, veľmi, *veľmi* opatrní predtým než začnete predpokladať, že je to takto, a začnete používať svoju moc nad druhými „aby ste ich rozhýbali“. Šéfovia, ktorí dokážu zistiť, kedy je niečo *naozaj* v rámci vašich schopností, ak ste o trochu viac motivovaný, bez toho aby ste vyhoreli, alebo aby sa váš život stal nesmierne bolestivým – to sú šéfovia, pod ktorými je radosť pracovať. *Táto schopnosť je extrémne zriedkavá*, a šéfovia, ktorí ju majú, sú hodní svojej váhy v striebre. Je to interpersonálna technika na vysokej úrovni, ktorú väčšina ľudí nemá. Ja ju určite nemám. Nepredpokladajte, že ju máte, pretože vaše úmysly sú dobré. Nepredpokladajte, že ju máte, pretože by ste *druhým* nikdy neurobili niečo, čo by ste nechceli, aby sa urobilo *vám*. Nepredpokladajte, že ju máte, pretože sa vám zatiaľ nikto nesťažoval. Možno sa jednoducho boja. Ten racionalista, o ktorom som hovoril – ktorý si nemyslel, že má moc a hrozby, hoci mne to určite bolo dosť zrejmé – on si neuvedomoval, že by sa ho niekto mohol báť.

Buďte opatrní, keď máte *páku*, keď máte vo svojich rukách dôležité rozhodnutie alebo hrozbu alebo niečo, čo ten druhý človek potrebuje, a náhle vám pokušenie optimalizovať ho pripadá neodolateľné.

Predstavte si, ak to dokážete, že celú vládu strachu Ayn Rand nad Objektivistami možno vidieť v tomto svetle – že zistila, že má moc a páku, a nedokázala odolať pokušeniu optimalizovať.

Podceňujeme vzdialenosť medzi nami a druhými. Nie iba inferenčnú vzdialenosť, ale vzdialenosť v povahe a schopnostiach, vzdialenosť v situácii a prostriedkoch, vzdialenosť v nevyslovených poznatkoch a nepovšimnutých zručnostiach a šťastí, vzdialenosť vo vnútornej rovine.

Dokonca aj ja som často prekvapený, keď zistím, že X, ktoré pre mňa tak dobre fungovalo, nefunguje pre niekoho iného. Ale keď sa ma tak veľa druhých ľudí pokúšalo optimalizovať, dokážem aspoň rozoznať vzdialenosť, keď ma udrie po hlave.

Možno byť tlačený do práce funguje... pre vás. Možno vás nezačne bolieť žalúdok, keď niekto, kto má nad vami moc, začne nápomocne skúšať preorganizovať váš život tým správnym spôsobom. Ja neviem, čo vás motivuje. V oblasti sily vôle a akrázie a produktivity, rovnako ako v iných oblastiach, nepoznám zovšeobecnenia dosť hlboké na to, aby platili takmer vždy. Nemám hlboké kľúče, ktoré by mi povedali, *kedy* a *prečo* a *pre koho* nejaká technika funguje alebo nefunguje. Všetko, čo môžem robiť, je byť pripravený to prijať, keď mi niekto povie, že to nefunguje... a pokračovať v hľadaní hlbších

zovšeobecnení, ktoré budú platiť všade, hlbších zákonov, ktorými sa riadia aj pravidlá aj výnimky, ktoré čakajú, až ich jedného dňa nájdeme.

*→

331. Praktické rady podopreté hlbokými teóriami

Kde bolo, tam bolo, Seth Roberts šiel na dovolenku do Európy a zistil, že začal chudnúť zatiaľ čo popíjal kalorické ovocné šťavy nezvyčajnej chuti.

Predstavte si, že by Roberts nevedel, a nikdy by sa nedozvedel, nič o metabolických rovnovážnych bodoch alebo asociáciách medzi chuťou a kalóriami – všetok tento sofistikovaný vedecký experimentálny výskum, ktorý sa robil na potkanoch a občas aj na ľuďoch.

Napísal by na svoj blog: „Aha, všetci! Mali by ste skúsiť tieto úžasné ovocné šťavy, ktoré spôsobujú, že chudnem!“ A to by bolo všetko. Pár ľudí by to vyskúšalo, *dočasne* by im to fungovalo (dokiaľ by asociácia medzi chuťou a kalóriami nenaskočila) a nikdy by neexistovala diéta Shangri-La *ako taká*.

Existujúca diéta Shangri-La je viditeľne neúplná – pre niektorých ľudí, ako som ja, vyzerá, že nefunguje, a neexistuje na to žiaden zrejmý dôvod, ani logika, ktorá by to dovoľovala. Ale dôvod, prečo z toho mnoho ľudí malo úžitok – dôvod, prečo to bol viac než iba jeden článok na blogu opisujúci trik, ktorý asi fungoval pre jedného človeka a pre nikoho iného – bol ten, že Roberts *poznal experimentálnu vedu, ktorá mu umožnila interpretovať to, čo videl, v pojmoch hlbokých činiteľov, ktoré naozaj existujú.*

Jedna rada na *Overcoming Bias / Less Wrong*, ktorú ľudia často uvádzali ako tú najdôležitejšia vec, ktorú sa tam naučili, bola myšlienka „spodného riadku“ - že keď raz vo svojej mysli napíšete záver, buď je vtedy už pravdivý alebo už nepravdivý, už múdry alebo už hlúpy, a žiadne množstvo neskoršej argumentácie to nemôže zmeniť, jedine ak by zmenilo záver. A toto sa priamo spája s ďalšou často spomínanou najdôležitejšou vecou, čo je myšlienka „strojov na poznanie“, mysle ako mapujúceho stroja, ktorý potrebuje indície ako palivo.

Keby som bol napísal iba jeden článok na blog, ktorý by hovoril: „Viete, naozaj by ste mali byť otvorenejší voči zmene názoru – je to veľmi dôležité – ach, áno, mali by ste aj venovať pozornosť indíciám.“ Toto by nebolo také užitočné. Nie iba preto, že by to bolo *menej presvedčivé*, ale pretože tie skutočné operácie by boli omnoho menej jasné bez explicitnej teórie za tým. Napríklad, čo je to *indícia*? Je to hocičo, čo znie ako silný argument? Mať explicitnú teóriu pravdepodobnosti a explicitné kauzálne vysvetlenie, čo robí uvažovanie efektívnym, robí *veľký* rozdiel v sile a podrobnostiach implementácie tej starej rady: „Maj otvorenú myseľ a venuj pozornosť indíciám.“

Dôležité je aj uvedomiť si, že *kauzálne teórie* sú omnoho častejšie pravdivé, ak ich máte z vedeckej učebnice než keď si ich len tak vymyslíte – je veľmi ľahké vymýšľať kognitívne štruktúry, ktoré vyzerajú ako kauzálne teórie, ale ani len neovládajú očakávanie, tobôž že by boli pravdivé.

Toto je ten rukopis, ktorý chcem sprostredkovať pomocou všetkých tých článkov, ktoré prepájajú experimenty kognitívnej vedy a teóriu pravdepodobnosti a epistemológiu s praktickými radami – že praktická rada sa v skutočnosti stane prakticky silnejšou, ak pôjdete a prečítate si o experimentoch kognitívnej vedy, alebo o teórii pravdepodobnosti, alebo dokonca o materialistickej epistemológii, a *uvedomíte si, čo vidíte*. Toto je to znamenie, ktoré môže odlíšiť *Less Wrong* od desaťtisícov iných blogov tvrdiacich, že ponúkajú rady.

Mohol by som vám povedať: „Viete, nakoľko ste spokojní vo svojím jedlom, pravdepodobne závisí viac od kvality jedla, než od toho, koľko ho zjete.“ A vy by ste si to prečítali a zabudli by ste na to, a impulz dokončiť celý tanier by vám stále pripadal rovnako silný. Ale ak vám poviem o necitlivosti voči rozsahu, o zanedbávaní trvania, a o pravidle „v najlepšom treba prestať“, zrazu si uvedomíte veľmi konkrétnym spôsobom, pri pohľade na svoj tanier, že si v spätnom pohľade vytvoríte takmer rovnakú

→ http://lesswrong.com/lw/9v/beware_of_otheroptimizing/

spomienku, či je vaša porcia veľká alebo malá; teraz máte hlbokú teóriu o pravidlách, ktorými sa *riadi* vaša pamäť, a teraz viete, že toto je to, čo tie pravidlá hovoria. (Takisto viete, že si máte zákusok odložiť na koniec.)

Chcem počuť, ako môžem prekonať akráziu – ako môže mať väčšiu silu vôle, alebo urobiť viac s menšou mentálnou bolesťou. Lenže existuje desaťtisíc ľudí tvrdiacich, že vedia v tomto poradiť, a vo väčšine prípadov je to na úrovni toho alternatívneho Setha Robertsa, ktorý iba hovorí ľuďom o úžasných účinkoch pitia ovocnej šťavy. Alebo v skutočnosti ešte horšie než to – sú to ľudia, ktorí sa snažia opísať vnútorné myšlienkové páky, ktoré potiahli, pre ktoré neexistujú štandardné slová, a na ktoré v skutočnosti nevedia ukázať. Pozrite si aj ilúziu transparentnosti, inferenčnú vzdialenosť, a dvojitú ilúziu transparentnosti. (Všimnite si, ako sa: „Preceňujete, koľko ste vysvetlili, a vaši poslucháči preceňujú, koľko počuli“ stane *omnoho silnejšou* radou, keď ju podopriem experimentom z kognitívnej vedy a trochou evolučnej psychológie?)

Myslím si, že rada, ktorú *potrebujem*, je od niekoho, kto prečítal celý kopec experimentálnej psychológie týkajúcej sa sily vôle, myšlienkových konfliktov, vyčerpania ega, prevrátenia preferencií, hyperbolického diskontovania, zrútenia self, pikoekonomiky, a tak ďalej, a komu sa v procese prekonávania svojej vlastnej akrázie podarilo pochopiť, čo urobil, v *naozaj všeobecnom zmysle* – vďaka experimentom, ktoré mu dali slovník kognitívnych javov, ktoré *naozaj existujú*, na rozdiel od javov, ktoré si len vymyslel. A navyše, niekto, kto dokáže *vysvetliť*, čo urobil, niekomu inému, opäť vďaka experimentálnemu a teoretickému slovníku, ktorý mu dovolí ukázať na replikovateľné experimenty, ktoré tieto myšlienky zhmotňujú vo veľmi konkrétnych výsledkoch alebo v matematicky čistých myšlienkach.

Všimnite si stupeň rastúcej náročnosti pri citovaní:

- *Konkrétne experimentálne výsledky* (na ktoré človeku stačí len prečítať si odborný článok, ideálne taký, ktorý hlásil $p < 0,01$, pretože $p < 0,05$ sa nemusí replikovať)
- *Kauzálne vysvetlenia, ktoré sú naozaj pravdivé* (ktoré možno spoľahlivo získať hľadaním teórií, ktoré používa väčšina v rámci danej vedy)
- *Správne interpretovaná matematika* (ohľadom ktorej mám problém niečo užitočné poradiť, pretože väčšina môjho vlastného matematického talentu je intuícia, ktorá sa nakopne skôr než mám šancu pouvažovať)

Ak neviete, komu dôverovať, alebo ak nedôverujete sami sebe, mali by ste sa na začiatku sústrediť na experimentálne výsledky, potom prejsť na rozmýšľanie pomocou kauzálnych teórií, ktoré sa bežne používajú v rámci vedy, a namáčať si prsty do matematiky a epistemológie s extrémnou opatrnosťou.

Ale praktické rady sa naozaj, naozaj *stanú* omnoho mocnejšími, keď za nimi stoja *konkrétne experimentálne výsledky, naozaj pravdivé kauzálne vysvetlenia, a správne interpretovaná matematika.*

* →

332. Hriech nedostatku sebadôvery

Existujú tri veľké hriechy zakorenené najmä v racionalistoch, a tretím z nich je nedostatok sebadôvery. Michael Vassar ma pravidelne obviňuje z tohto hriechu, čo ho robí jedinečným v celej populácii tejto Zeme.

Ale v skutočnosti má celkom pravdu, keď sa tým trápi, aj ja sa tým trápim, a každý šikovný racionalista asi strávi značné množstvo času trápením sa s tým. Keď pokusné osoby vedia o nejakom skreslení alebo sú na nejaké skreslenie upozornené, *prehnaná oprava* nie je medzi pokusnými výsledkami ničím neznámym. Preto je veľa kognitívnych podúloh takých náročných – viete, že ste skreslení, ale nie ste si istí, *nakoľko*, a neviete, či sa opravujete *dosť* – a tak by ste sa možno mali opraviť trochu viac, a potom ešte trochu viac, ale je *to* dosť? Alebo ste to azda ďaleko prestrelili? Ste na tom teraz možno ešte horšie, než keby ste sa vôbec nepokúšali o žiadnu opravu?

→ http://lesswrong.com/lw/d4/practical_advice_backed_by_deep_theories/

Rozjímate nad tou vecou, cítite sa viac a viac stratení, a samotná úloha odhadovať vám začne pripadať čoraz márnejšia...

A keď dôjde na konkrétne otázky *sebadôvery*, *prehnanej sebadôvery*, a *nedostatočnej sebadôvery* – chápané teraz v širšom zmysle slova, nie iba ako kalibrované intervaly spoľahlivosti – existuje prirodzený sklon vyzdvihnúť prehnanú sebadôveru ako *hriech* pýchy, z toho *druhého* zoznamu, ktorý nás nikdy nevaroval pred nesprávnym použitím pokory alebo zneužívaním pochybnosti. Umiestniť sa príliš vysoko – vyskakovať nad svoje primerané miesto – myslieť si o sebe priveľa – tlačiť sa dopredu – implicitným porovnaním zhadzovať svojich blížnych – a dôsledky poníženia a zvrhnutia, možno verejného – nie sú to hnusné a hrozné veci?

Byť príliš *skromný*... vyzerá v porovnaní s tým menej vážne; nebolo by až také ponižujúce, keby vám to verejne vytkli, veru, zistenie, že ste lepší než ste si predstavovali, by mohlo prísť ako hrejivé prekvapenie; klásť seba nízko, a druhých implicitne vyššie, má v sebe pozitívnych nádych *prívetivosti*, je to niečo také, ako by urobil Gandalf.

Ak ste sa teda naučili tisíc spôsobov, ako sa ľudia mýlia, a prečítali stovky experimentálnych výsledkov, v ktorých sú anonymné pokusné osoby ponížené vďaka svojmu prehnanému sebavedomiu – sakra, ak ste ich prečítali čo len pár tuctov – a *neviete*, nakoľko presne prehnané sebavedomie máte vy – potom áno, môžete byť naozaj v nebezpečenstve, že sa posuniete príliš hlboko dole.

Neviem vám dať žiaden dokonalý vzorec, ktorý by tomuto zabránil. Ale mám radu alebo dve.

V čom je *riziko* nedostatku sebadôvery?

Odmietanie príležitostí. Nerobenie vecí, ktoré ste mohli urobiť, ale ste sa nepokúsili (dosť silno).

Tu je teda prvá rada: Ak existuje spôsob, ako *zistiť*, nakoľko ste dobrí, treba to *otestovať*. *Hypotéza si zaslúži testovanie*; platí to aj pre hypotézy o vašich vlastných schopnostiach. Jedného dňa mi pripadalo, že by som mal dokázať vyhrať v experimente s UI v krabici; a vyzeralo to ako veľmi pochybná a arogantná myšlienka; tak som to otestoval. Neskôr mi pripadalo, že by som mal dokázať vyhrať dokonca aj keď budú v stávke veľké peniaze, otestoval som to, ale vyhral som iba v 1 prípade z 3. To bola teda v danom čase hranica mojej schopnosti, a nepotreboval som argumentovať smerom nahor ani nadol, pretože som to mohol skrátka *otestovať*.

Jeden z hlavných spôsobov, ako bystrí ľudia skončia ako hlupáci, je natoľko si *zvyknúť na vyhrávanie*, že sa pridržiavajú miest, kde *vedia, že môžu vyhrať* – čo znamená, že nikdy nenatiahnu svoje schopnosti, nikdy sa nepokúsia o nič zložité.

Hovorí sa, že toto súvisí s tým, či sami seba definujete pomocou „inteligencie“ viac než „úsilia“, pretože vtedy je *ľahké* vyhrávanie príznakom vašej „inteligencie“, zatiaľ čo zlyhanie pri ťažkom probléme by sa stále dalo interpretovať ako dobré úsilie.

Som si teda celkom istý, že zdatný racionalista by mal o týchto veciach rozmýšľať takto: rozumnosť *je* systematizované vyhrávanie, a pokúsiť sa pokúsiť vyzerá ako cesta k zlyhaniu. Povedal by som to takto: Hypotéza si zaslúži testovanie! Ak *neviete*, či vyhráte pri ťažkom probléme – potom *potrápte svoju rozumnosť*, aby ste *objavili* svoju terajšiu úroveň. Nezvyknem schvaľovať, keď si ľudia blahoželajú za to, že sa pokúsili – to mi pripadá ako zlý myšlienkový zvyk – ale *nepokúsiť* sa je iste ešte *horšie*. Ak si pestujete všeobecný zvyk púšťať sa do problémov, a vyhráte aspoň pri *niektorých*, potom si snáď môžete pomyslieť: „Držal som sa svojho zvyku púšťať sa do problémov, a budem tak robiť aj nabudúce.“ Môžete si aj pomyslieť: „Získal som hodnotnú informáciu o svojej terajšej úrovni a kde sa potrebujem zlepšiť,“ pokiaľ túto myšlienku správne doplníte o: „nebudem sa pokúšať zistiť tú istú hodnotnú informáciu aj nabudúce“.

Ak *vždy* vyhráte, znamená to, že sa dostatočne nevyvíjate. Ale *mali* by ste sa zakaždým vážne snažiť vyhrať. A ak sa budete príliš utešovať za zlyhanie, stratíte svojho súťaživého ducha a zakrpatiete.

Keď si skúsim predstaviť, čo by na toto povedal fiktívny majster Kompetitívnej Konšpirácie, vychádza mi niečo ako: „Nie je *dobré* prehrávať. Ale *bolesť* z prehry nie je niečo také strašné, aby ste utekali pred výzvou pretože sa jej bojíte. Nie je to také strašné, aby ste sa tomuto pocitu museli opatrne vyhýbať, alebo odmietať pripustiť, že ste prehrali, a to poriadne. Prehra *má* bolieť. Keby vás nebolela,

neboli by ste Kompetitor. Neexistuje *žiaden* Kompetitor, ktorý *nikdy* nepoznal bolesť z prehry. A teraz choďte a *vyhrajte*."

Pestujte si zvyk púšťania sa do problémov – azda nie takých, ktoré vás rovno zabijú, ale azda takých, ktoré vás môžu *ponížiť*. Nedávno som čítal o istom veriacom, že porazil v debate Christophera Hitchensa (dosť drvivo; to povedali ateisti). A tak som napísal ľuďom z Bloggingheads a opýtal som sa, či by mi mohli dohodnúť debatu. Vyzeralo to ako niekto, s kým by som sa rád otestoval. Tiež sa hovorilo, že Christopher Hitchens si mal pozrieť predchádzajúce debaty daného veriaceho a byť pripravený, tak som sa rozhodol, že to *neurobím*, pretože si myslím, že by som mal dokázať zvládnuť takmer všetko naživo, a chcem sa dozvedieť, či je táto myšlienka pravdivá; a som ochotný riskovať verejné poníženie, aby som to zistil. Poznamenávam, že toto *nie je* sebahendikepovanie v klasickom zmysle – ak sa táto debata naozaj dohodne (zatiaľ nemám odpoveď) a ja sa nepripravím a zlyhám, potom prehrám v tej stávke, ktorú som si stanovil; získam informáciu o mojich hraniciach; *nedal* som si nič, čo by som považoval za výhovorku v prípade prehry.

Samozrejme takto možno uvažovať iba keď *naozaj* čelíte problému len preto, aby ste sa otestovali, a nie preto, lebo musíte za každú cenu vyhrať. V *takom* prípade si treba všetko zjednodušiť, ako sa len dá. Konať inak by bolo *okázalá* prehnaná sebadôvera, ešte aj keby ste hrali piškvorky proti trojročnému dieťaťu.

Jemnejšia forma nedostatočnej sebadôvery je *strácať svoju zotrvačnosť v pohybe vpred* – medzi všetkými tými vecami, ktoré si uvedomíte, že ľudia robia nesprávne, ktoré ste aj vy robili nesprávne, ktoré pravdepodobne stále robíte nesprávne. Začnete byť bojazlivý; začnete pochybovať o sebe, ale *neodpoviete na tieto pochybnosti a nepôjdete ďalej*; keď si vytvoríte hypotézu o svojej neschopnosti, *nepodrobíte túto hypotézu testu*.

Možno ani nebude viditeľný nejaký smutný okamih, keď sa vedome, viditeľne vo vlastných očiach *rozhodnete neskúsiť* nejaký konkrétny test,... iba... akosi... spomalíte...

Zdá sa, že sa už neoplatí ísť a pokúšať sa opraviť jednu vec, keď existuje tucet ďalších vecí, ktoré môžu byť stále zle...

Nezostáva dosť nádeje na triumf, aby vás *inšpirovala tvrdo* pracovať...

Keď sa zamyslíte nad robením hocičoho nového, okamžite vám do hlavy naskočí tucet otázok o vašej schopnosti, a nenapadne vám, že by ste na tieto otázky mohli *odpovedať* tým, že sa *otestujete*...

A keďže ste čítali tak veľa múdrosti o ľudských chybách, začne sa vám zdať, že múdrosť spočíva vo večnom pochybovaní (nikdy nie v riešení pochybností), vo večnej pokore odmietania (nikdy nie v pokore prípravy), a že tak všeobecne je múdre hovoriť o ľudských schopnostiach horšie a horšie veci, prepracovať sa k cynickému dobrému pocitu zo zlého pocitu.

A tak moja posledná rada je iný uhol pohľadu, z ktorého sa môžete pozrieť na daný problém – z ktorého môže posudzovať všetky možné myšlienkové zvyky, ktoré by ste si mohli osvojiť – a to je pýtať sa:

Robí ma tento spôsob myslenia silnejším alebo slabším? Skutočne naozaj?

Už som hovoril o nebezpečenstve *rozumného znenia* – rozumne znejúci argument, že by sme v Newcombovom probléme mali vziať dve krabice, rozumne znejúci argument, že nemôžeme nič vedeť vďaka problému indukcie, rozumne znejúci argument, že by sme na tom boli v priemere lepšie, keby sme vždy prijali názor väčšiny, a ďalšie takéto prekážky v Ceste. „Vyhráva to?" je otázka, ktorú si môžete položiť, aby ste získali alternatívny pohľad. Iný, mierne odlišný pohľad je opýtať sa: „Robí ma tento spôsob myslenia silnejším alebo slabším?" Robí vás ustavičné si pripomínanie, že máte o všetom pochybovať, silnejším alebo slabším? Robí vás odmietanie vyriešiť alebo zmenšiť tieto pochybnosti silnejším alebo slabším? Robí vás podstúpenie vedomej krízy viery zoči-voči neistote silnejším alebo slabším? Robí vás odpovedanie na každú námietku pokorným vyznaním vašej omylnosti silnejším alebo slabším?

Robia vás vaše terajšie pokusy vyvažovať možnú prehnanú sebadôveru silnejším alebo slabším? Nápoveda: Ak sa viac pripravujete, svedomitejšie sa testujete, žiadate priateľov o radu, prepracovávate sa k veľkým veciam postupne, alebo stále občas zlyhávate, ale menej ako kedysi, potom sa pravdepodobne

stávate silnejším. Ak *nikdy* nezlyhávate, vyhýbate sa výzvam, a vo všeobecnosti sa cítite beznádejne a skleslo, potom sa pravdepodobne stávate slabším.

Prvú formu tohto pravidla som sa naučil vo veľmi mladom veku, keď som si cvičil istý matematický test, a zistil som, že po každom cvičnom teste moje skóre klesalo, a keď som si prešiel hárok s odpoveďami, všimol som si, že som tie správne odpovede mal napísané ceruzkou a potom vygumované. Tak som si povedal: „Dobre, *tentokrát* použijem Silu a bude konať podľa inštinktu," a moje skóre vyletelo na ešte vyššie než som mal na začiatku, a na skutočnom teste bolo ešte vyššie. Tak som sa naučil, že pochybovať o sebe vás nerobí vždy silnejšími – najmä ak to prekáža vašej schopnosti nechať sa rozhýbať dobrými informáciami, ako sú napríklad vaše matematické intuície. (Ale *potreboval* som test, aby mi to povedal!)

Nedostatok sebadôvery nie je hriech jedinečný pre racionalistov. Ale je to konkrétne nebezpečenstvo, do ktorého vás môže doviesť *úsilie byť rozumný*. A je to *fatálna* chyba – chyba, ktorá vám bráni získať ďalšiu skúsenosť, ktorá by túto chybu mohla opraviť.

Pretože nedostatok sebavedomia v skutočnosti *vyzerá* byť pomerne častý medzi ašpirujúcimi racionalistami, ktorých som stretol – hoci pomerne zriedkavejší medzi racionalistami, z ktorých sa stali slávne vzory – menujem ho ako tretí medzi troma hriechmi zakorenenými v racionalistoch.

* →

333. Choďte ďalej a vytvorte umenie!

Počas posledných mesiacov som povedal pár vecí o rozumnosti. Povedal som pár vecí o tom, ako rozmotať otázky, ktoré sa stali zmätené, a ako rozoznať rozdiel medzi skutočným rozmýšľaním a falošným rozmýšľaním, a o vôli stať sa silnejším, ktorá vás vedie skúšať skôr než ujdete; povedal som aj niečo o robení nemožného.

A toto všetko sú techniky, ktoré som vyvinul počas svojich vlastných projektov – čo je dôvod, prečo je tam napríklad toľko o kognitívnom redukcionizme – a je možné, že keď ich skúsite použiť na seba, budete potrebovať niečo iné. Môžete potrebovať niečo iné. Napriek tomu, tí, ktorí sa pýtajú: „Ale načo je to dobré?" by si mohli skúsiť opäť prečítať niektoré zo skorších článkov; vedieť napríklad o klame konjunkcie a ako si zbadať v argumente, sotva vyzerá ezotericky. Rozumieť, prečo vám motivovaný skepticizmus škodí, môže podľa mňa zodpovedať za celý rozdiel medzi bystrým človekom, ktorý skončí ako bystrý, a bystrým človekom, ktorý skončí ako hlúpy. Afektívne špirály smrti pohltili *mnohých*, ktorí si nedávali pozor...

Napriek tomu si myslím, že v tomto „umení rozumnosti" stále viac *chýba* než tam *je* – ako poraziť akráziu a ako koordinovať skupiny, to sú dve veci z tých, ktorých nedostatok ma najviac trápi. Vo všeobecnosti som sa viac sústredil na epistemickú rozumnosť než na inštrumentálnu rozumnosť. A potom je tu cvičenie, vyučovanie, overovanie, a vytvorenie z toho poriadnej experimentálnej vedy. A ak to trochu ďalej zovšeobecníte, potom by *vytvorenie Umenia* mohlo zahŕňať aj veci ako vytvorenie lepšej úvodnej literatúry, vytvorenie lepších sloganov pre styk s verejnosťou, zriadenie spoločnej kauzy s ďalšími podúlohami Osvietenstva, analýza a riešenie problému nerovnováhy medzi pohlaviami...

Ale tie malé kúsky rozumnosti, ktoré som pripravil... *dúfam*... že možno...

Mám podozrenie – mohli by ste to nazvať aj odhad – že existuje nejaká *bariéra pri štarte*, v týchto veciach rozumnosti. Kde štandardne, na začiatku, nemáte na čom budovať. Máte toho až tak málo, že nemáte ani len náznak, že existuje viac, že existuje nejaké Umenie, ktoré možno nájsť. A ak začnete cítiť, že sa dá aj viac – potom je možné, že *okamžite* nesprávne odbočíte. Ako hovorí David Stove, väčšina „veľkých mysliteľov" vo filozofii, napríklad Hegel, si v skutočnosti zaslúži ľútosť.[301] Takéto veci sa štandardne stávajú každému, kto sa podujme vybudovať umenie myslenia; vyvinie falošné odpovede.

→ http://lesswrong.com/lw/c3/the_sin_of_underconfidence/

301 David Charles Stove, *The Plato Cult and Other Philosophical Follies* (Cambridge University Press, 1991).

Keď sa snažíte vyvinúť časť umenia ľudského myslenia... potom robíte niečo *nie celkom nepodobné* tomu, čo som robil v oblasti umelej inteligencie. Budú vás pokúšať falošné vysvetlenia mysle, falošné výklady kauzality, tajomné sväté slová, a nejaká úžasná myšlienka, ktorá vyrieši všetko.

Nie je to tak, že by tie konkrétne, epistemické metódy odhaľovania falošnosti, ktoré používam, boli také dobré na každý *konkrétny* problém; ale zdá sa, že by mohli byť užitočné pri rozlišovaní dobrých a zlých *spôsobov myslenia*.

Dúfam, že niekto, kto sa naučí tú časť Umenia, ktorú som tu predložil, neodbočí *okamžite* a *automaticky* nesprávnym smerom, keď sa opýta sám seba: „Ako by ľudia mali myslieť, aby dokázali vyriešiť nový problém X, na ktorom pracujem?“ Neujde okamžite preč; nebude si len tak vymýšľať náhodné veci; možno ho to pohne k štúdiu literatúry v experimentálnej psychológii; neupadne automaticky do afektívnej špirály smrti ohľadom svojej Skvelej Myšlienky; bude mať nejakú predstavu, čím sa líši falošné vysvetlenie od skutočného. Dostane hod proti pasci.

Možno je to tento druh bariéry, ktorý bráni ľuďom *začať* vyvíjať umenie rozumnosti, ak ešte nie sú rozumní.

A tak namiesto toho... idú mimo a vymyslia Freudovskú psychoanalýzu. Alebo nové náboženstvo. Alebo niečo. To je to, čo sa stáva *štandardne*, keď ľudia začnú rozmýšľať o rozmýšľaní.

Dúfam, že tá časť Umenia, ktorú som vytvoril, nech je akokoľvek neúplná, dokáže prekonať túto úvodnú bariéru – dať ľuďom základ, na ktorom sa dá stavať; dať im predstavu, že Umenie existuje, a niečo o tom, ako by sa malo vyvíjať; a dať im prinajmenšom *hod proti pasci* predtým než *okamžite* nesprávne odbočia.

To je môj sen – že toto napohľad vysoko špecializované umenie odpovedania na zmätené otázky môže byť tým, čo treba, na samotnom začiatku, aby sa dalo *ísť a doplniť zvyšok*.

Úloha, ktorú nechávam na *vás*. Teda, pravdepodobne. Nesľubujem nič ohľadom toho, kam sa moja pozornosť môže obrátiť v budúcnosti. Ale, viete, *sú* nejaké iné veci, ktoré potrebujem urobiť. Dokonca aj keby som náhodou vyvinul ešte viac Umenia, možno už nebudem mať čas o ničom z toho písať.

Okrem toho všetkého, čo už som povedal o falošných odpovediach a pasciach, sú tu dve veci, ktoré by som bol rád, keby ste na ne pamätali.

Po prvé – že som pri tvorbe svojho Umenia čerpal z viacerých zdrojov. Čítal som mnohých rôznych autorov, mnoho rôznych experimentov, používal analógie z mnohých rôznych oblastí. Aj *vy* budete musieť čerpať z viacerých zdrojov, aby ste vytvorili *svoju* časť Umenia. Nemali by ste dostať všetku svoju rozumnosť od jedného autora – hoci by azda mohla existovať nejaká centralizovaná webová stránka, kam pôjdete uverejniť odkazy a články, ktoré vám pripadali naozaj dôležité. Zrejúce Umenie bude musieť čerpať z viacerých zdrojov. Podľa môjho najlepšieho vedomia neexistuje *žiadna* skutočná veda, ktorá čerpá svoju silu iba od jednej osoby. Podľa môjho najlepšieho vedomia, takýmto spôsobom fungujú *výhradne* sekty. Skutočná veda môže mať svojich hrdinov, môže mať dokonca svojich osamelých rebelských hrdinov, ale *bude mať viac než jedného*.

Po druhé – že som vytvoril svoje Umenie v procese *skúšania urobiť nejakú konkrétnu vec*, ktorá motivovala všetko moje úsilie. Možno som príliš idealistický – možno priveľa rozmýšľam o tom, ako by svet *mal* fungovať – ale aj tak mám isté podozrenie, že nemôžete vyvinúť Umenie *iba* tým, že budete sedieť a rozmýšľať: „Ako by som mohol bojovať proti tejto večičke akrázii?“ Zvyšok Umenia môžete vyvinúť v procese snahy *urobiť niečo*. Možno dokonca – ak teraz prehnane nezovšeobecňujem z mojej vlastnej histórie – nejakú úlohu dosť zložitú na to, aby ponamáhala a zlomila vaše staré chápanie a donútila vás znovu vymyslieť pár vecí. Ale možno sa mýlim, a ďalšia časť práce bude urobená priamym, konkrétnym skúmaním „rozumnosti“, bez akejkoľvek potreby konkrétneho využitia, ktoré by bolo považované za dôležitejšie.

Moje predchádzajúce pokusy opísať tento princíp pomocou úcty ohraničenej tajnou totožnosťou moje obecenstvo zborovo odmietlo. Mohlo by „odísť z domu“ byť vhodnejšie? *Pripadá* mi to ako naozaj dobrá, zdravá myšlienka. Ale – možno sa mýlim. Uvidíme, odkiaľ naozaj prídu ďalšie časti Umenia.

Dlho som sa usiloval sprostredkovať, odovzdať, podeliť sa o časť tej zvláštnej veci, ktorej som sa dotkol, ktorá mi pripadá taká vzácna. A nie som si istý, že som niekedy vyjadril tento ústredný rytmus

slovami. Možno ho dokážete nájsť počúvaním tónov. Viem povedať tieto slová, ale nie to pravidlo, ktoré ich vytvorilo, alebo pravidlo za týmto pravidlom; človek môže iba dúfať, že *použitím* tých myšlienok sa azda vo vás môže zrodiť podobné zariadenie. Pamätajte si, že *všetko* ľudské úsilie o naučenie sa tajomstiev štandardne skĺzalo do hesiel, chválospevov, a vznášajúcich sa tvrdení.

Dlho som sa usiloval sprostredkovať svoje Umenie. Väčšinou bez úspechu, pred týmto terajším pokusom. Predtým som sa snažil iba občas, a možno som dostal toľko úspechu, koľko som si zaslúžil. Ako keď hádžete do jazera kamienky, ktoré vytvoria pár vĺn, a potom vymiznú... Tentokrát som do toho dal veľa driny a zdvihol som veľkú skalu. Čas ukáže, či bola dosť veľká – či som naozaj niekoho *rozrušil* dosť hlboko, aby tie vlny po dopade pokračovali svojím vlastným pohybom. Čas ukáže, či som vytvoril niečo, čo sa hýbe svojou vlastnou silou.

Chcem, aby ľudia išli ďalej, ale aj aby sa vrátili. Alebo možno, aby zároveň aj išli ďalej, aj zostali, pretože toto je internet a tu si takúto vec môžeme dovoliť; v poslednej dobe som sa na *Less Wrong* naučil pár zaujímavých vecí, a ak je udržanie si motivácie celé roky problémom, rozprávanie sa s druhými (alebo len videnie, že ostatní sa tiež snažia) často pomáha.

Ale každopádne, ak som vás nejako ovplyvnil, potom dúfam, že pôjdete ďalej a pustíte sa do problémov, a dosiahnete niekam mimo svojho kresla, a vytvoríte nové Umenie; a potom, keď si spomeniete, odkiaľ ste prišli, odkážete vysielačkou ostatným, čo ste sa naučili.

* →

Literatúra

Albert, David Z. Quantum Mechanics and Experience. Harvard University Press, 1994.

Alexander, Scott. „Why I Am Not Rene Descartes.“ Slate Star Codex (blog) (2014). http://slatestarcodex.com/2014/11/27/why- i-am-not- renedescartes/.

Allais, Maurice. „Le Comportement de l’Homme Rationnel devant le Risque: Critique des Postulats et Axiomes de l’Ecole Americaine.“ Econometrica 21, no. 4 (1953): 503–546. doi:10.2307/1907921.

Ariely, Dan. Predictably Irrational: The Hidden Forces That Shape Our Decisions. HarperCollins, 2008.

Asch, Solomon E. „Studies of Independence and Conformity: A Minority of One Against a Unanimous Majority.“ Psychological Monographs 70 (1956).

Ashmore, Richard D., Vasantha Ramchandra, and Russell A. Jones. „Censorship as an Attitude Change Induction.“ Paper presented at Eastern Psychological Association meeting (1971).

Asimov, Isaac. The Relativity of Wrong. Oxford University Press, 1989.

Baez, John. „The Crackpot Index.“ 1998. http://math.ucr.edu/home/baez/ crackpot.html.

Banks, Iain. The Player of Games. Orbit, 1989.

Baratz, Daphna. How Justified Is the „Obvious“ Reaction? Stanford University, 1983.

Barbour, Julian. The End of Time: The Next Revolution in Physics. 1st ed. New York: Oxford University Press, 1999.

Baron, Jonathan. Thinking and Deciding. Cambridge University Press, 2007.

Baron, Jonathan, and Joshua D. Greene. „Determinants of Insensitivity to Quantity in Valuation of Public Goods: Contribution, Warm Glow, Budget Constraints, Availability, and Prominence.“ Journal of Experimental Psychology: Applied 2, no. 2 (1996): 107–125. doi:10.1037/1076-898X.2.2.107.

Barrett, Jeffrey. Everett’s Relative-State Formulation of Quantum Mechanics. Edited by Edward N. Zalta. http://plato.stanford.edu/archives/fall2008/ entries/qm-everett/.

Benson, Peter L., Stuart A. Karabenick, and Richard M. Lerner. „Pretty Pleases: The Effects of Physical Attractiveness, Race, and Sex on Receiving Help.“ Journal of Experimental Social Psychology 12 (5 1976): 409– 415. doi:10.1016/0022-1031(76)90073-1.

→ http://lesswrong.com/lw/c4/go_forth_and_create_the_art/

Bond, Rod, and Peter B. Smith. „Culture and Conformity: A Meta-Analysis of Studies Using Asch's (1952b, 1956) Line Judgment Task.“ Psychological Bulletin 119 (1996): 111–137.

Bostrom, Nick. „A History of Transhumanist Thought.“ Journal of Evolution and Technology 14, no. 1 (2005): 1–25. http://www.nickbostrom.com/ papers/history.pdf.

. „Existential Risks: Analyzing Human Extinction Scenarios and Related Hazards.“ Journal of Evolution and Technology 9 (2002). http://www. jetpress.org/volume9/risks.html.

Bostrom, Nick. Superintelligence: Paths, Dangers, Strategies. Oxford University Press, 2014.

Bostrom, Nick, and Milan M. Ćirković, eds. Global Catastrophic Risks. New York: Oxford University Press, 2008.

Bostrom, Nick, and Julian Savulescu. „Human Enhancement Ethics: The State of the Debate.“ In Human Enhancement, edited by Nick Bostrom and Julian Savulescu. 2009.

Bostrom, Nick, and Eliezer Yudkowsky. „The Ethics of Artificial Intelligence.“ In The Cambridge Handbook of Artificial Intelligence, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press, 2014.

Bourget, David, and David J. Chalmers. „What Do Philosophers Believe?“ Philosophical Studies (2013): 1–36.

Boynton, Robert S. „The Birth of an Idea: A Profile of Frank Sulloway.“ The New Yorker (October 1999).

Brehm, Sharon S., and Marsha Weintraub. „Physical Barriers and Psychological Reactance: Two-year-olds' Responses to Threats to Freedom.“ Journal of Personality and Social Psychology 35 (1977): 830–836.

Broeder, Dale. „The University of Chicago Jury Project.“ Nebraska Law Review 38 (1959): 760–774.

Brown, Kevin. Reflections On Relativity. Raleigh, NC: printed by author, 2011. http://www.mathpages.com/rr/rrtoc.htm.

Brust, Steven. The Paths of the Dead. Vol. 1 of The Viscount of Adrilankha. Tor Books, 2002.

Budesheim, Thomas Lee, and Stephen DePaola. „Beauty or the Beast?: The Effects of Appearance, Personality, and Issue Information on Evaluations of Political Candidates.“ Personality and Social Psychology Bulletin 20 (4 1994): 339–348. doi:10.1177/0146167294204001.

Buehler, Roger, Dale Griffin, and Michael Ross. „Exploring the ‚Planning Fallacy': Why People Underestimate Their Task Completion Times.“ Journal of Personality and Social Psychology 67, no. 3 (1994): 366–381. doi:10.1037/0022-3514.67.3.366.

. „Inside the Planning Fallacy: The Causes and Consequences of Optimistic Time Predictions.“ In Gilovich, Griffin, and Kahneman, Heuristics and Biases, 250–270.

. „It's About Time: Optimistic Predictions in Work and Love.“ European Review of Social Psychology 6, no. 1 (1995): 1–32. doi:10.1080/14792779343000112.

Bujold, Lois McMaster. Komarr. Miles Vorkosigan Adventures. Baen, 1999.

Burton, Ian, Robert W. Kates, and Gilbert F. White. The Environment as Hazard. 1st ed. New York: Oxford University Press, 1978.

Campbell, Richmond, and Lanning Snowden, eds. Paradoxes of Rationality and Cooperation: Prisoner's Dilemma and Newcomb's Problem. Vancouver: University of British Columbia Press, 1985.

Carson, Richard T., and Robert Cameron Mitchell. „Sequencing and Nesting in Contingent Valuation Surveys.“ Journal of Environmental Economics and Management 28, no. 2 (1995): 155–173. doi:10.1006/jeem.1995.1011.

Casscells, Ward, Arno Schoenberger, and Thomas Graboys. „Interpretation by Physicians of Clinical Laboratory Results.“ New England Journal of Medicine 299 (1978): 999–1001.

Castellow, Wilbur A., Karl L. Wuensch, and Charles H. Moore. „Effects of Physical Attractiveness of the Plaintiff and Defendant in Sexual Harassment Judgments.“ Journal of Social Behavior and Personality 5 (6 1990): 547– 562.

Chaiken, Shelly. „Communicator Physical Attractiveness and Persuasion.“ Journal of Personality and Social Psychology 37 (8 1979): 1387–1397. doi:10.1037/0022-3514.37.8.1387.

Chalmers, David J. The Conscious Mind: In Search of a Fundamental Theory. New York: Oxford University Press, 1996.

Chapman, Graham, et al. Monty Python’s The Life of Brian (of Nazareth). Eyre Methuen, 1979.

Chapman, Gretchen B., and Eric J. Johnson. „Incorporating the Irrelevant: Anchors in Judgments of Belief and Value.“ In Gilovich, Griffin, and Kahneman, Heuristics and Biases, 120–138.

Cherryh, Carolyn J. The Paladin. Baen, 2002.

Cialdini, Robert B. Influence: Science and Practice. Boston: Allyn & Bacon, 2001.

. Influence: The Psychology of Persuasion: Revised Edition. New York: Quill, 1993.

Cleaver, Jerry. Immediate Fiction: A Complete Writing Course. Macmillan, 2004.

Combs, Barbara, and Paul Slovic. „Newspaper Coverage of Causes of Death.“ Journalism & Mass Communication Quarterly 56, no. 4 (1979): 837–849. doi:10.1177/107769907905600420.

Crawford, Charles B., Brenda E. Salter, and Kerry L. Jang. „Human Grief: Is Its Intensity Related to the Reproductive Value of the Deceased?“ Ethology and Sociobiology 10, no. 4 (1989): 297–307.

Darwin, Charles. On the Origin of Species by Means of Natural Selection; or, The Preservation of Favoured Races in the Struggle for Life. 1st ed. London: John Murray, 1859. http://darwin-online.org.uk/content/frameset? viewtype=text&itemID=F373&pageseq=1.

. The Descent of Man, and Selection in Relation to Sex. 2nd ed. London: John Murray, 1874. http://darwin- online.org.uk/content/frameset? itemID=F944&viewtype=text&pageseq=1.

Darwin, Francis, ed. The Life and Letters of Charles Darwin. Vol. 2. John Murray, 1887.

Dawes, Robyn M. House of Cards: Psychology and Psychotherapy Built on Myth. Free Press, 1996.

. Rational Choice in An Uncertain World. 1st ed. Edited by Jerome Kagan. San Diego, CA: Harcourt Brace Jovanovich, 1988.

De Camp, Lyon Sprague, and Fletcher Pratt. The Incomplete Enchanter. New York: Henry Holt & Company, 1941.

Denes-Raj, Veronika, and Seymour Epstein. „Conflict between Intuitive and Rational Processing: When People Behave against Their Better Judgment.“ Journal of Personality and Social Psychology 66 (5 1994): 819–829. doi:10.1037/0022-3514.66.5.819.

Dennett, Daniel C. Breaking the Spell: Religion as a Natural Phenomenon. Penguin, 2006.

. Darwin’s Dangerous Idea: Evolution and the Meanings of Life. Simon & Schuster, 1995.

. Freedom Evolves. Viking Books, 2003.

. „The Unimagined Preposterousness of Zombies.“ Journal of Consciousness Studies 2 (4 1995): 322–26.

Descartes, René. Discours de la Méthode. Vol. 45. Librairie des Bibliophiles, 1887.

Desvousges, William H., F. Reed Johnson, Richard W. Dunford, Kevin J. Boyle, Sara P. Hudson, and K. Nicole Wilson. Measuring Nonuse Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy. Technical report. Research Triangle Park, NC: RTI International, 2010. doi:10.3768/rtipress.2009.bk.0001.1009.

Downs, A. Chris, and Phillip M. Lyons. „Natural Observations of the Links Between Attractiveness and Initial Legal Judgments.“ Personality and Social Psychology Bulletin 17 (5 1991): 541–547. doi:10.1177/0146167291175009.

Drescher, Gary L. Good and Real: Demystifying Paradoxes from Physics to Ethics. Cambridge, MA: MIT Press, 2006.

Eagly, Alice H., Richard D. Ashmore, Mona G. Makhijani, and Laura C. Longo. „What Is Beautiful Is Good, But . . . A Meta-analytic Review of Research on the Physical Attractiveness Stereotype.“ Psychological Bulletin 110 (1 1991): 109–128. doi:10.1037/0033-2909.110.1.109.

Eddy, David M. „Probabilistic Reasoning in Clinical Medicine: Problems and Opportunities.“ In Judgement Under Uncertainty: Heuristics and Biases, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky. Cambridge University Press, 1982.

Efran, M. G., and E. W. J. Patterson. „The Politics of Appearance.“ Unpublished PhD thesis, 1976.

Egan, Greg. Quarantine. London: Legend Press, 1992.

Ehrlinger, Joyce, Thomas Gilovich, and Lee Ross. „Peering Into the Bias Blind Spot: People's Assessments of Bias in Themselves and Others.“ Personality and Social Psychology Bulletin 31, no. 5 (2005): 680–692.

Eidelman, Scott, and Christian S. Crandall. „Bias in Favor of the Status Quo.“ Social and Personality Psychology Compass 6, no. 3 (2012): 270–281.

Epley, Nicholas, and Thomas Gilovich. „Putting Adjustment Back in the Anchoring and Adjustment Heuristic: Differential Processing of Self- Generated and Experimentor-Provided Anchors.“ Psychological Science 12 (5 2001): 391–396. doi:10.1111/1467-9280.00372.

Epstein, Lewis Carroll. Thinking Physics: Understandable Practical Reality, 3rd Edition. Insight Press, 2009.

Feldman, Richard. „Naturalized Epistemology.“ In The Stanford Encyclopedia of Philosophy, Summer 2012, edited by Edward N. Zalta.

Festinger, Leon, Henry W. Riecken, and Stanley Schachter. When Prophecy Fails: A Social and Psychological Study of a Modern Group That Predicted the Destruction of the World. Harper-Torchbooks, 1956.

Fetherstonhaugh, David, Paul Slovic, Stephen M. Johnson, and James Friedrich. „Insensitivity to the Value of Human Life: A Study of Psychophysical Numbing.“ Journal of Risk and Uncertainty 14, no. 3 (1997): 283–300. doi:10.1023/A:1007744326393.

Feynman, Richard P. „Judging Books by Their Covers.“ In Surely You're Joking, Mr. Feynman! New York: W. W. Norton & Company, 1985. Feynman, Richard P., Robert B. Leighton, and Matthew L. Sands. The Feynman Lectures on Physics. 3 vols. Reading, MA: Addison-Wesley, 1963.

Finucane, Melissa L., Ali Alhakami, Paul Slovic, and Stephen M. Johnson. „The Affect Heuristic in Judgments of Risks and Benefits.“ Journal of Behavioral Decision Making 13, no. 1 (2000): 1–17.

Fong, Geoffrey T., David H. Krantz, and Richard E. Nisbett. „The Effects of Statistical Training on Thinking about Everyday Problems.“ Cognitive Psychology 18, no. 3 (1986): 253–292. doi:10.1016/0010-0285(86)90001-0.

Frank, Adam. The Constant Fire: Beyond the Science vs. Religion Debate. University of California Press, 2009.

Freire, Paulo. The Politics of Education: Culture, Power, and Liberation. Greenwood Publishing Group, 1985.

Galbraith, John Kenneth. Economics, Peace and Laughter. Plume, 1981.

Ganzach, Yoav. „Judging Risk and Return of Financial Assets.“ Organizational Behavior and Human Decision Processes 83, no. 2 (2000): 353– 370. doi:10.1006/obhd.2000.2914.

Gendlin, Eugene T. Focusing. Bantam Books, 1982.

Gibbon, Edward. The History of the Decline and Fall of the Roman Empire. Vol. 4. J. & J. Harper, 1829.

Gigerenzer, Gerd, and Ulrich Hoffrage. „How to Improve Bayesian Reasoning without Instruction: Frequency Formats.“ Psychological Review 102 (1995): 684–704.

Gilbert, Daniel T., Douglas S. Krull, and Patrick S. Malone. „Unbelieving the Unbelievable: Some Problems in the Rejection of False Information.“ Journal of Personality and Social Psychology 59 (4 1990): 601–613. doi:10.1037/0022-3514.59.4.601.

Gilbert, Daniel T., and Patrick S. Malone. „The Correspondence Bias.“ Psychological Bulletin 117, no. 1 (1995): 21–38. http://www.wjh.harvard.edu/ ~dtg/Gilbert%20&%20Malone %20(CORRESPONDENCE%20BIAS) .pdf.

Gilbert, Daniel T., Romin W. Tafarodi, and Patrick S. Malone. „You Can't Not Believe Everything You Read.“ Journal of Personality and Social Psychology 65 (2 1993): 221–233. doi:10.1037/0022-3514.65.2.221.

Gilbert, William S., and Arthur Sullivan. The Mikado. Opera, 1885.

Gilovich, Thomas, Dale Griffin, and Daniel Kahneman, eds. Heuristics and Biases: The Psychology of Intuitive Judgment. New York: Cambridge University Press, 2002. doi:10.2277/0521796792.

Goertzel, Ben, and Cassio Pennachin, eds. Artificial General Intelligence. Cognitive Technologies. Berlin: Springer, 2007. doi:10.1007/978-3-540-68677-4.

Good, Irving John. „Speculations Concerning the First Ultraintelligent Machine." In Advances in Computers, edited by Franz L. Alt and Morris Rubinoff, 6:31–88. New York: Academic Press, 1965. doi:10.1016/S0065-2458(08)60418-0.

Graur, Dan, and Wen-Hsiung Li. Fundamentals of Molecular Evolution. 2nd ed. Sunderland, MA: Sinauer Associates, 2000.

Griffin, Dale, and Amos Tversky. „The Weighing of Evidence and the Determinants of Confidence." Cognitive Psychology 24, no. 3 (1992): 411–435. doi:10.1016/0010-0285(92)90013-R.

Haidt, Jonathan. „The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." Psychological Review 108, no. 4 (2001): 814–834. doi:10.1037/0033-295X.108.4.814.

Halberstadt, Jamin Brett, and Gary M. Levine. „Effects of Reasons Analysis on the Accuracy of Predicting Basketball Games." Journal of Applied Social Psychology 29, no. 3 (1999): 517–530.

Haldane, John B. S. „A Mathematical Theory of Natural and Artificial Selection." Mathematical Proceedings of the Cambridge Philosophical Society 23 (5 1927): 607–615. doi:10.1017/S0305004100011750.

Hamermesh, Daniel S., and Jeff E. Biddle. „Beauty and the Labor Market." The American Economic Review 84 (5 1994): 1174–1194.

Hansen, Katherine, Margaret Gerbasi, Alexander Todorov, Elliott Kruse, and Emily Pronin. „People Claim Objectivity After Knowingly Using Biased Strategies." Personality and Social Psychology Bulletin 40, no. 6 (2014): 691–699.

Hanson, Robin. „You Are Never Entitled to Your Opinion." Overcoming Bias (blog) (2006). http://www.overcomingbias.com/2006/12/you_are_never_ e.html.

Harris, Sam. The End of Faith: Religion, Terror, and the Future of Reason. WW Norton & Company, 2005.

Hastorf, Albert, and Hadley Cantril. „They Saw a Game: A Case Study." Journal of Abnormal and Social Psychology 49 (1954): 129–134. http://www2. psych.ubc.ca/~schaller/Psyc590Readings/Hastorf1954.pdf.

Heuer, Richards J. Psychology of Intelligence Analysis. Center for the Study of Intelligence, Central Intelligence Agency, 1999.

Hibbard, Bill. „Super-Intelligent Machines." ACM SIGGRAPH Computer Graphics 35, no. 1 (2001): 13–15. http://www.siggraph.org/publications/ newsletter/issues/v35/v35n1.pdf.

Hintze, Daniel. „Problem Class Dominance in Predictive Dilemmas." Honors thesis (2014).

Hodgell, Patricia C. Seeker's Mask. Meisha Merlin Publishing, Inc., 2001.

Holt, Jim. „Thinking Inside the Boxes." Slate (2002). http://www.slate.com/ articles/arts/egghead/2002/02/thinkinginside%5C_the%5C_boxes.single. html.

Holyoak, Keith J., and Robert G. Morrison. The Oxford Handbook of Thinking and Reasoning. Oxford University Press, 2013.

Hsee, Christopher K. „Less Is Better: When Low-Value Options Are Valued More Highly than High-Value Options." Behavioral Decision Making 11 (2 1998): 107–121.

. „The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives." Organizational Behavior and Human Decision Processes 67 (3 1996): 247–257. doi:10.1006/obhd.1996.0077.

Hsee, Christopher K., and Howard C. Kunreuther. „The Affection Effect in Insurance Decisions." Journal of Risk and Uncertainty 20 (2 2000): 141– 159. doi:10.1023/A:1007876907268.

Hutter, Marcus. Universal Artificial Intelligence: Sequential Decisions Based On Algorithmic Probability. Berlin: Springer, 2005. doi:10.1007/b138233.

Huxley, Thomas Henry. Evolution and Ethics and Other Essays. Macmillan, 1894. Imagination Engines, Inc. „The Imagination Engine® or Imagitron™.“ 2011. http://www.imagination-engines.com/ie.htm.

Jaynes, Edwin T. Probability Theory: The Logic of Science. Edited by George Larry Bretthorst. New York: Cambridge University Press, 2003. doi:10.2277/0521592712.

. „Probability Theory as Logic.“ In Maximum Entropy and Bayesian Methods, edited by Paul F. Fougère. Springer Netherlands, 1990.

. „Probability Theory, with Applications in Science and Engineering.“ Unpublished manuscript (1974).

Jones, Edward E., and Victor A. Harris. „The Attribution of Attitudes.“ Journal of Experimental Social Psychology 3 (1967): 1–24. http://www.radford. edu/~jaspelme/443/spring-2007/Articles/Jones_n_Harris_1967.pdf.

Joyce, James M. The Foundations of Causal Decision Theory. New York: Cambridge University Press, 1999. doi:10.1017/CBO9780511498497.

Kahneman, Daniel. „Comments by Professor Daniel Kahneman.“ In Valuing Environmental Goods: An Assessment of the Contingent Valuation Method, edited by Ronald G. Cummings, David S. Brookshire, and William D. Schulze, vol. 1.B, 226–235. Experimental Methods for Assessing Environmental Benefits. Totowa, NJ: Rowman & Allanheld, 1986. http://yosemite. epa.gov/ee/epa/eerm.nsf/vwAN/EE-0280B-04.pdf/$file/EE-0280B- 04.pdf.

Kahneman, Daniel, Ilana Ritov, and David Schkade. „Economic Preferences or Attitude Expressions?: An Analysis of Dollar Responses to Public Issues.“ Journal of Risk and Uncertainty 19, nos. 1–3 (1999): 203–235. doi:10.1023/A:1007835629236.

Kahneman, Daniel, David A. Schkade, and Cass R. Sunstein. „Shared Outrage and Erratic Awards: The Psychology of Punitive Damages.“ Journal of Risk and Uncertainty 16 (1 1998): 48–86. doi:10.1023/A:1007710408413.

Kahneman, Daniel, and Amos Tversky. „Prospect Theory: An Analysis of Decision Under Risk.“ Econometrica 47 (1979): 263–292.

Keats, John. „Lamia.“ The Poetical Works of John Keats (London: Macmillan) (1884).

Keysar, Boaz. „Language Users as Problem Solvers: Just What Ambiguity Problem Do They Solve?“ In Social and Cognitive Approaches to Interpersonal Communication, edited by Susan R. Fussell and Roger J. Kreuz, 175–200. Mahwah, NJ: Lawrence Erlbaum Associates, 1998.

. „The Illusory Transparency of Intention: Linguistic Perspective Taking in Text.“ Cognitive Psychology 26 (2 1994): 165–208. doi:10.1006/cogp.1994.1006.

Keysar, Boaz, and Dale J. Barr. „Self-Anchoring in Conversation: Why Language Users Do Not Do What They ‚Should.“‘ In Gilovich, Griffin, and

Kahneman, Heuristics and Biases, 150–166.

Keysar, Boaz, and Bridget Bly. „Intuitions of the Transparency of Idioms: Can One Keep a Secret by Spilling the Beans?“ Journal of Memory and Language 34 (1 1995): 89–109. doi:10.1006/jmla.1995.1005.

Keysar, Boaz, and Anne S. Henly. „Speakers’ Overestimation of Their Effectiveness.“ Psychological Science 13 (3 2002): 207–212. doi:10.1111/1467-9280.00439.

Kirk, Robert. Mind and Body. McGill-Queen’s University Press, 2003.

Knishinsky, A. „The Effects of Scarcity of Material and Exclusivity of Information on Industrial Buyer Perceived Risk in Provoking a Purchase Decision.“ Doctoral dissertation, Arizona State University, 1982.

Kosslyn, Stephen M. Image and Brain: The Resolution of the Imagery Debate. Cambridge, MA: MIT Press, 1994.

Krips, Henry. „Measurement in Quantum Theory.“ In The Stanford Encyclopedia of Philosophy, Fall 2013, edited by Edward N. Zalta.

Kruschke, John K. Doing Bayesian Data Analysis, Second Edition: A Tutorial with R, JAGS, and Stan. Academic Press, 2014.

. „What to Believe: Bayesian Methods for Data Analysis.“ Trends in Cognitive Sciences 14, no. 7 (2010): 293–300.

Kulka, Richard A., and Joan B. Kessler. „Is Justice Really Blind?: The Effect of Litigant Physical Attractiveness on Judicial Judgment.“ Journal of Applied Social Psychology 8 (4 1978): 366–381. doi:10.1111/j.1559-1816.1978.tb00790.x.

Kunreuther, Howard, Robin Hogarth, and Jacqueline Meszaros. „Insurer Ambiguity and Market Failure.“ Journal of Risk and Uncertainty 7 (1 1993): 71–87. doi:10.1007/BF01065315.

Lakoff, George. Women, Fire, and Dangerous Things: What Categories Reveal about the Mind. Chicago: Chicago University Press, 1987.

Latané, Bibb, and John M. Darley. „Bystander ‚Apathy.“‘ American Scientist 57, no. 2 (1969): 244–268. http://www.jstor.org/stable/27828530.

Lazarsfeld, Paul F. „The American Solidier—An Expository Review.“ Public Opinion Quarterly 13, no. 3 (1949): 377–404.

Le Guin, Ursula K. The Farthest Shore. Saga Press, 2001.

Ledwig, Marion. „Newcomb’s Problem.“ PhD diss., University of Constance, 2000.

Lichtenstein, Sarah, and Paul Slovic. „Reversals of Preference Between Bids and Choices in Gambing Decisions.“ Journal of Experimental Psychology 89, no. 1 (1971): 46–55. , eds. The Construction of Preference. Cambridge University Press, 2006.

Lichtenstein, Sarah, Paul Slovic, Baruch Fischhoff, Mark Layman, and Barbara Combs. „Judged Frequency of Lethal Events.“ Journal of Experimental Psychology: Human Learning and Memory 4, no. 6 (1978): 551–578. doi:10.1037/0278-7393.4.6.551.

Liersch, Michael J., and Craig R. M. McKenzie. „Duration Neglect by Numbers and Its Elimination by Graphs.“ Organizational Behavior and Human Decision Processes 108, no. 2 (2009): 303–314.

Mack, Denise, and David Rainey. „Female Applicants’ Grooming and Personnel Selection.“ Journal of Social Behavior and Personality 5 (5 1990): 399–407.

MacKay, David J. C. Information Theory, Inference, and Learning Algorithms. New York: Cambridge University Press, 2003.

Matthews, Robert. „Do We Need to Change the Definition of Science?“ New Scientist (May 2008).

Mazis, Michael B. „Antipollution Measures and Psychological Reactance Theory: A Field Experiment.“ Journal of Personality and Social Psychology 31 (1975): 654–666.

Mazis, Michael B., Robert B. Settle, and Dennis C. Leslie. „Elimination of Phosphate Detergents and Psychological Reactance.“ Journal of Marketing Research 10 (1973): 390–395.

McDermott, Drew. „Artificial Intelligence Meets Natural Stupidity.“ SIGART Newsletter, no. 57 (1976): 4–9. doi:10.1145/1045339.1045340.

McFadden, Daniel L., and Gregory K. Leonard. „Issues in the Contingent Valuation of Environmental Goods: Methodologies for Data Collection and Analysis.“ In Contingent Valuation: A Critical Assessment, edited by Jerry A. Hausman, 165–215. Contributions to Economic Analysis 220. New York: North-Holland, 1993. doi:10.1108/S0573-8555(1993)0000220007.

Mercier, Hugo, and Dan Sperber. „Why Do Humans Reason? Arguments for an Argumentative Theory.“ Behavioral and Brain Sciences 34 (2011): 57– 74. https://hal.archives-ouvertes.fr/file/index/docid/904097/filename/ MercierSperberWhydohumansreason.pdf.

Meyers, David G. Exploring Social Psychology. New York: McGraw-Hill, 1994.

Mitchell, Tom M. Machine Learning. McGraw-Hill Science/Engineering/Math, 1997.

Moore, John. Slay and Rescue. Xlibris Corp, 2000.

Muehlhauser, Luke. „The Power of Agency.“ Less Wrong (blog) (2011). http: //lesswrong.com/lw/5i8/the_power_of_agency/.

Musashi, Miyamoto. Book of Five Rings. New Line Publishing, 2003.

Mussweiler, Thomas, and Fritz Strack. „Comparing Is Believing: A Selective Accessibility Model of Judgmental Anchoring.“ European Review of Social Psychology 10 (1 1999): 135–167. doi:10.1080/14792779943000044.

Nagel, Thomas. „What Is It Like to Be a Bat?“ Philosophical Review 83, no. 4 (1974): 435–450. http://www.jstor.org/stable/2183914.

Newby-Clark, Ian R., Michael Ross, Roger Buehler, Derek J. Koehler, and Dale Griffin. „People Focus on Optimistic Scenarios and Disregard Pessimistic Scenarios While Predicting Task Completion Times.“ Journal of Experimental Psychology: Applied 6, no. 3 (2000): 171–182. doi:10.1037/1076-898X.6.3.171.

Nickerson, Raymond S. „Confirmation Bias: A Ubiquitous Phenomenon in Many Guises.“ Review of General Psychology 2, no. 2 (1998): 175.

Nisbett, Richard E., and Timothy D. Wilson. „Telling More than We Can Know: Verbal Reports on Mental Processes.“ Psychological Review 84 (1977): 231–259. http://people.virginia.edu/~tdw/nisbett&wilson.pdf.

Orwell, George. 1984. Signet Classic, 1950.

. „Politics and the English Language.“ Horizon (April 1946).

. „Why Socialists Don’t Believe in Fun.“ Tribune (December 1943).

Pearce, David. The Hedonistic Imperative. http://www.hedweb.com/, 1995.

Pearl, Judea. Causality: Models, Reasoning, and Inference. 2nd ed. New York: Cambridge University Press, 2009.

. Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. San Mateo, CA: Morgan Kaufmann, 1988.

Peirce, Charles Sanders. Collected Papers. Harvard University Press, 1931.

Pinker, Steven. The Blank Slate: The Modern Denial of Human Nature. New York: Viking, 2002.

Piper, Henry Beam. „Police Operation.“ Astounding Science Fiction (July 1948).

Pirsig, Robert M. Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values. 1st ed. New York: Morrow, 1974.

Planck, Max. Scientific Autobiography and Other Papers. New York: Philosophical Library, 1949.

Plato. Great Dialogues of Plato. Edited by Eric H. Warmington and Philip G. Rouse. Signet Classic, 1999.

Popper, Karl R. The Logic of Scientific Discovery. New York: Basic Books, 1959.

Poundstone, William. Priceless: The Myth of Fair Value (and How to Take Advantage of It). Hill & Wang, 2010.

Pratchett, Terry. Maskerade. Discworld Series. ISIS, 1997.

. Witches Abroad. London: Corgi Books, 1992. Procopius. History of the Wars. Edited by Henry B. Dewing. Vol. 1. Harvard University Press, 1914.

Pronin, Emily. „How We See Ourselves and How We See Others.“ Science 320 (2008): 1177–1180. http://psych.princeton.edu/psychology/research/ pronin/pubs/2008%20Self%20and%20Other.pdf.

Pronin, Emily, Daniel Y. Lin, and Lee Ross. „The Bias Blind Spot: Perceptions of Bias in Self versus Others.“ Personality and Social Psychology Bulletin 28, no. 3 (2002): 369–381.

Putnam, Hilary. „The Meaning of Meaning.“ In The Twin Earth Chronicles, edited by Andrew Pessin and Sanford Goldberg, 3–52. M. E. Sharpe, Inc., 1996.

Quattrone, George A., Cheryl P. Lawrence, Steven E. Finkel, and David C. Andrus. „Explorations in Anchoring: The Effects of Prior Range, Anchor Extremity, and Suggestive Hints.“ Unpublished manuscript, Stanford University, 1981.

Rhodes, Richard. The Making of the Atomic Bomb. New York: Simon & Schuster, 1986.

Rips, Lance J. „Inductive Judgments about Natural Categories.“ Journal of Verbal Learning and Verbal Behavior 14 (1975): 665–681.

Roberts, Seth. „What Makes Food Fattening?: A Pavlovian Theory of Weight Control.“ Unpublished manuscript, 2005. http://media.sethroberts.net/ about/whatmakesfoodfattening.pdf.

Robinson, Dave, and Judy Groves. Philosophy for Beginners. 1st ed. Cambridge: Icon Books, 1998.

Rorty, Richard. „Out of the Matrix: How the Late Philosopher Donald Davidson Showed That Reality Can't Be an Illusion." The Boston Globe (October 2003).

Rosch, Eleanor. „Principles of Categorization." In Cognition and Categorization, edited by Eleanor Rosch and Barbara B. Lloyd. Hillsdale, NJ: Lawrence Erlbaum, 1978.

Russell, Gillian. „Epistemic Viciousness in the Martial Arts." In Martial Arts and Philosophy: Beating and Nothingness, edited by Graham Priest and Damon A. Young. Open Court, 2010.

Russell, Stuart J., and Peter Norvig. Artificial Intelligence: A Modern Approach. 3rd ed. Upper Saddle River, NJ: Prentice-Hall, 2010.

Ryle, Gilbert. The Concept of Mind. University of Chicago Press, 1949.

Sadock, Jerrold. „Truth and Approximations." Papers from the Third Annual Meeting of the Berkeley Linguistics Society (1977): 430–439.

Sagan, Carl. The Demon-Haunted World: Science as a Candle in the Dark. 1st ed. New York: Random House, 1995.

Saint-Exupery, Antoine de. Terre des Hommes. Paris: Gallimard, 1939.

Sandel, Michael. „What's Wrong With Enhancement." Background material for the President's Council on Bioethics. (2002).

Schoemaker, Paul J. H. „The Role of Statistical Knowledge in Gambling Decisions: Moment vs. Risk Dimension Approaches." Organizational Behavior and Human Performance 24, no. 1 (1979): 1–17.

Schul, Yaacov, and Ruth Mayo. „Searching for Certainty in an Uncertain World: The Difficulty of Giving Up the Experiential for the Rational Mode of Thinking." Journal of Behavioral Decision Making 16, no. 2 (2003): 93– 106. doi:10.1002/bdm.434.

Schwitzgebel, Eric. Perplexities of Consciousness. MIT Press, 2011. Serfas, Sebastian. Cognitive Biases in the Capital Investment Context: Theoretical Considerations and Empirical Experiments on Violations of Normative Rationality. Springer, 2010.

Sherif, Muzafer, Oliver J. Harvey, B. Jack White, William R. Hood, and Carolyn W. Sherif. „Study of Positive and Negative Intergroup Attitudes Between Experimentally Produced Groups: Robbers Cave Study." Unpublished manuscript (1954).

Shermer, Michael. „The Unlikeliest Cult in History." Skeptic 2, no. 2 (1993): 74–81. http://www.2think.org/02_2_she.shtml.

Shimabukuro, Masayuki. Flashing Steel: Mastering Eishin-Ryu Swordsmanship. Frog Books, 1995.

Slovic, Paul. „Numbed by Numbers." Foreign Policy (March 2007). http:// foreignpolicy.com/2007/03/13/numbed-by-numbers/.

Slovic, Paul, Melissa Finucane, Ellen Peters, and Donald G. MacGregor. „Rational Actors or Rational Fools: Implications of the Affect Heuristic for Behavioral Economics." Journal of Socio-Economics 31, no. 4 (2002): 329– 342. doi:10.1016/S1053-5357(02)00174-9.

Smith, Edward Elmer. Second Stage Lensmen. Old Earth Books, 1998.

Smith, Edward Elmer, and Ric Binkley. Gray Lensman. Old Earth Books, 1998.

Smith, Edward Elmer, and A. J. Donnell. First Lensman. Old Earth Books, 1997.

Smullyan, Raymond M. What Is the Name of This Book?: The Riddle of Dracula and Other Logical Puzzles. Penguin Books, 1990.

Soares, Nate, and Benja Fallenstein. „Toward Idealized Decision Theory." Technical report. Berkeley, CA: Machine Intelligence Research Institute (2014). http://intelligence.org/files/TowardIdealizedDecisionTheory.pdf.

Spee, Friedrich. Cautio Criminalis; or, A Book on Witch Trials. Edited and translated by Marcus Hellyer. Studies in Early Modern German History. 1631. Charlottesville: University of Virginia Press, 2003.

Stanovich, Keith E., and Richard F. West. „Individual Differences in Reasoning: Implications for the Rationality Debate?" Behavioral and Brain Sciences 23, no. 5 (2000): 645–665. http://journals.cambridge.org/abstract_ S0140525X00003435.

Stewart, John E. „Defendants' Attractiveness as a Factor in the Outcome of Trials: An Observational Study." Journal of Applied Social Psychology 10 (4 1980): 348–361. doi:10.1111/j.1559-1816.1980.tb00715.x.

Stiegler, Marc. David's Sling. Baen, 1988.

Stove, David Charles. The Plato Cult and Other Philosophical Follies. Cambridge University Press, 1991.

Strack, Fritz, and Thomas Mussweiler. „Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility." Journal of Personality and Social Psychology 73, no. 3 (1997): 437–446.

Sunstein, Cass R. „Probability Neglect: Emotions, Worst Cases, and Law." Yale Law Journal (2002): 61–107.

Taber, Charles S., and Milton Lodge. „Motivated Skepticism in the Evaluation of Political Beliefs." American Journal of Political Science 50, no. 3 (2006): 755–769. doi:10.1111/j.1540-5907.2006.00214.x.

Talbott, William. „Bayesian Epistemology." In The Stanford Encyclopedia of Philosophy, Fall 2013, edited by Edward N. Zalta.

TalkOrigins Foundation. „Frequently Asked Questions about Creationism and Evolution." http://www.talkorigins.org/origins/faqs-qa.html.

Tegmark, Max. Our Mathematical Universe: My Quest for the Ultimate Nature of Reality. Random House LLC, 2014.

. „Parallel Universes." In Science and Ultimate Reality: Quantum Theory, Cosmology, and Complexity, edited by John D. Barrow, Paul C. W. Davies, and Charles L. Harper Jr., 459–491. New York: Cambridge University Press, 2004.

Thompson, Silvanus Phillips. The Life of Lord Kelvin. American Mathematical Society, 2005.

Thornton, Stephen. „Karl Popper." In The Stanford Encyclopedia of Philosophy, Winter 2002, edited by Edward N. Zalta. Stanford University. http://plato. stanford.edu/archives/win2002/entries/popper/.

Tooby, John, and Leda Cosmides. „The Psychological Foundations of Culture." In The Adapted Mind: Evolutionary Psychology and the Generation of Culture, edited by Jerome H. Barkow, Leda Cosmides, and John Tooby, 19–136. New York: Oxford University Press, 1992.

Tversky, Amos, and Ward Edwards. „Information versus Reward in Binary Choices." Journal of Experimental Psychology 71, no. 5 (1966): 680–683. doi:10.1037/h0023123.

Tversky, Amos, and Itamar Gati. „Studies of Similarity." In Cognition and Categorization, edited by Eleanor Rosch and Barbara Lloyd, 79–98. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc., 1978.

Tversky, Amos, and Daniel Kahneman. „Extensional Versus Intuitive Reasoning: The Conjunction Fallacy in Probability Judgment." Psychological Review 90, no. 4 (1983): 293–315. doi:10.1037/0033-295X.90.4.293.

. „Judgment Under Uncertainty: Heuristics and Biases." Science 185, no. 4157 (1974): 1124–1131. doi:10.1126/science.185.4157.1124.

. „Judgments of and by Representativeness." In Judgment Under Uncertainty: Heuristics and Biases, edited by Daniel Kahneman, Paul Slovic, and Amos Tversky, 84–98. New York: Cambridge University Press, 1982.

Tzu, Sun. The Art of War. Cloud Hands, Inc., 2004.

Uhlmann, Eric Luis, and Geoffrey L. Cohen. „‚I think it, therefore it's true': Effects of Self-perceived Objectivity on Hiring Discrimination." Organizational Behavior and Human Decision Processes 104, no. 2 (2007): 207– 223.

Vaidman, Lev. „Many-Worlds Interpretation of Quantum Mechanics." In The Stanford Encyclopedia of Philosophy, Fall 2008, edited by Edward N. Zalta.

Vallone, Robert P., Lee Ross, and Mark R. Lepper. „The Hostile Media Phenomenon: Biased Perception and Perceptions of Media Bias in Coverage of the Beirut Massacre." Journal of Personality and Social Psychology 49 (1985): 577–585. http://ssc.wisc.edu/~jpiliavi/965/hwang.pdf.

Verhagen, Joachim. http : / / web . archive . org / web / 20060424082937 / http : //www.nvon.nl/scheik/best/diversen/scijokes/scijokes.txt. Archived version, October 27, 2001.

Wade, Michael J. „Group selections among laboratory populations of Tribolium.“ Proceedings of the National Academy of Sciences of the United States of America 73, no. 12 (1976): 4604–4607. doi:10.1073/pnas.73.12.4604.

Wansink, Brian, Robert J. Kent, and Stephen J. Hoch. „An Anchoring and Adjustment Model of Purchase Quantity Decisions.“ Journal of Marketing Research 35, no. 1 (1998): 71–81. http://www.jstor.org/stable/3151931.

Warren, Adam. Empowered. Vol. 1. Dark Horse Books, 2007.

Wason, Peter Cathcart. „On the Failure to Eliminate Hypotheses in a Conceptual Task.“ Quarterly Journal of Experimental Psychology 12, no. 3 (1960): 129–140. doi:10.1080/17470216008416717.

Weiten, Wayne. Psychology: Themes and Variations, Briefer Version, Eighth Edition. Cengage Learning, 2010.

West, Richard F., Russell J. Meserve, and Keith E. Stanovich. „Cognitive Sophistication Does Not Attenuate the Bias Blind Spot.“ Journal of Personality and Social Psychology 103, no. 3 (2012): 506.

Williams, George C. Adaptation and Natural Selection: A Critique of Some Current Evolutionary Thought. Princeton Science Library. Princeton, NJ: Princeton University Press, 1966.

Wilson, David Sloan. „A Theory of Group Selection.“ Proceedings of the National Academy of Sciences of the United States of America 72, no. 1 (1975): 143–146.

Wilson, Timothy D., David B. Centerbar, and Nancy Brekke. „Mental Contamination and the Debiasing Problem.“ In Heuristics and Biases: The Psychology of Intuitive Judgment, edited by Thomas Gilovich, Dale Griffin, and Daniel Kahneman. Cambridge University Press, 2002.

Wilson, Timothy D., Douglas J. Lisle, Jonathan W. Schooler, Sara D. Hodges, Kristen J. Klaaren, and Suzanne J. LaFleur. „Introspecting About Reasons Can Reduce Post-choice Satisfaction.“ Personality and Social Psychology Bulletin 19 (1993): 331–331.

Wittgenstein, Ludwig. Philosophical Investigations. Translated by Gertrude E. M. Anscombe. Oxford: Blackwell, 1953.

Wright, Robert. The Moral Animal: Why We Are the Way We Are: The New Science of Evolutionary Psychology. Pantheon Books, 1994.

Yamagishi, Kimihiko. „When a 12.86% Mortality Is More Dangerous than 24.14%: Implications for Risk Communication.“ Applied Cognitive Psychology 11 (6 1997): 461–554.

Yates, J. Frank, Ju-Whei Lee, Winston R. Sieck, Incheol Choi, and Paul C. Price. „Probability Judgment Across Cultures.“ In Gilovich, Griffin, and Kahneman, Heuristics and Biases, 271–291.

Yudkowsky, Eliezer. „Artificial Intelligence as a Positive and Negative Factor in Global Risk.“ In Bostrom and Ćirković, Global Catastrophic Risks, 308– 345.

. „Cognitive Biases Potentially Affecting Judgment of Global Risks.“ In Bostrom and Ćirković, Global Catastrophic Risks, 91–119.

. Timeless Decision Theory. Unpublished manuscript. Machine Intelligence Research Institute, Berkeley, CA, 2010. http://intelligence.org/files/ TDT.pdf.

Yuzan, Daidoji, William Scott Wilson, Jack Vaughn, and Gary Miller Haskins. Budoshoshinshu: The Warrior’s Primer of Daidoji Yuzan. Black Belt Communications Inc., 1984.

Zapato, Lyle. „Lord Kelvin Quotations.“ 2008. http://zapatopi.net/kelvin/ quotes/.

Zelazny, Roger. Prince of Chaos. Thorndike Press, 2001.

Zhuangzi and Thomas Merton. The Way of Chuang Tzu. New Directions Publishing, 1965.

Zhuangzi and Burton Watson. The Complete Works of Zhuangzi. Columbia University Press, 1968.