IN DEFENSE OF HYPER-LOGICIST AI

Selmer Bringsjord & Michael Zenzen
Department of Philosophy
Rensselaer Polytechnic Institute
Troy, New York 12180-3590
userctkw@rpitsmts.bitnet

Abstract

We argue for "hyper"-logicism, the view, hitherto unarticulated, that AI can succeed in creating a genuine robot agent by building a symbol system of the appropriate sort which has no sub-symbolic interaction whatsoever with the external world.

1 Introduction

A rather unfriendly debate continues to rage in AI between the "logicists" and the "connectionists." (Hereafter, the 'C-L' debate.) Many connectionists (e.g., (Smolensky 1988), (Churchland et al. 1990), (Waltz 1988), (Schwartz 1988), (Kaplan Weaver et al. 1990)) hold that their "brain-like" architectures ought to supplant or at least supplement the symbol-based ones of traditional logicist AI. On the other hand, many logicists (e.g. (Fodor and Pylyshyn 1988)) hold that any successful AI model of human cognition, and *a fortiori* any sentient artificial intelligence itself, must use classical, logic-driven architecture.

In this paper we argue for "hyper"-logicism, the most extreme form of logicism — an approach to AI hitherto unarticulated, let alone defended. In a word, hyper-logicism is the view that AI can succeed in creating a genuine robot agent by building a symbol system of the appropriate sort which has no sub-symbolic interaction whatsoever with the "external" world. What is pictured here is thus apparently a symbol system that has no need for processing and representation schemes near and dear to the heart of connectionists.

Two assumptions underlie the coming argumentation, and are worth setting out before we embark.

We assume, first, that the sort of AI with which we are concerned, whether it be connectionist or logicist or hybrid in spirit, is, at bottom, *aggressive*. Someone who views AI as nothing more than the attempt to do things like model computationally the olfactory component of rat brains will find the debate with which we are concerned to be otiose. On the other hand, if one has a sanguine, rounded view of AI, our treatment should be of interest. Such a view of AI, from our perspective, is two-fold in nature, namely that ATs engineering side is reflected by the aim of *building* a genuine artificial intelligence (= a mind, person, or agent — not necessarily of the human variety), while it's scientific side is reflected by the fact that reaching the engineering objective requires a thorough *understanding* of mentality itself.

Our second assumption is simply that readers of what is to come are to an appreciable degree familiar with the concepts central to the debate in question. We assume here, in particular, that readers have a background, on the connectionist side, largely derivable from volume I of *PDP* (Rumelhart and McClelland 1986), that is that they have assimilated neural net concepts like input, output and hidden units, activation, values, weights, training with back propagation, and so-called recurrent nets. On the logicist side, we assume readers to be at least comfortable with *n*-order extensional logics, timid to full-blown intensional logics, and traditional symbolic projects in AI employing fragments of these symbolic schemes. Furthermore we assume that the reader in in command of the basic concepts and proofs of elementary computability theory, e.g., finite state automata, Turing machines, K-tape Turing machines, cellular automata, and simulation proofs (e.g.) of the fact that a K-tape Turing machine is no more powerful than a standard one, that a cellular automaton can be viewed as just a K-tape Turing machine, and that a neural net can be recast as, among other things, a probabilistic cellular automaton. We would, in addition, like to assume that readers are familiar with analog devices, but this is perhaps unreasonable, since not only are physical analog computers in short supply, but also there is no satisfactory logico-mathematical or philosophic definition of 'X is an analog computer.'

It would be nice if readers had a formal understanding of symbol systems (sufficient, say, to assimilate Lindstroms First and Second Theorems — see Chapter XII of (Ebbinghaus, Flum et al. 1984)), but command of an informal account of the sort given by (Harnad 1990) suffices. Formally speaking, Hamad's account of a symbol system is erected by simply building on, and then generalizing from, the basic machinery of first-order logic, which, as is well-known, has both a "derivation" side, and a "meaning" side. That is, a symbol system is simply a generalization from, and perhaps if need be a refinement of, a first-order understanding of the familiar V and 'k' One generalizes from a first-order scheme by allowing symbol strings to be arbitrarily coded — as marks on paper, microevents in a digital computer or brain, etc, and also by stipulating that derivations are based, to use (Harnad 1990)'s term, merely on the *shape* of symbol strings.

2 Strong Logicism and Connectionism

The classic account of strong logicism is given by (Fodor and Pylyshyn 1988), who argue that connectionism may provide theories of the *implementation* of cognition, but not theories of psychology, i.e., not theories of what is *apparently* (emphasis to preclude begging any questions) symbolic in nature — things for example like the ability of a human person to produce/understand a sentence *S* if there is another related sentence 5' which this person produces/understands. (E.g., set *S* = 'John loves Mary,' set *S'* = 'Mary loves John.[1])

Strong logicism apparently includes the view that symbol systems will meet connectionist processing when these systems are "hooked up" to the external world. The view here is that connectionist work can proceed on its own, separate from logicist research, and then down the road, when the logicist wants to have her robot agent interact with the physical world, the fruit of connectionist research can be applied. Strong logicists, for example (Fodor 1980, 1985) have thus typically espoused such slogans as that 'their symbol system must be connected to the world in the right way/ They thus embrace a view of robot agents possessed of nonsymbolic "transducers" which would allow symbolic processing at the heart of the robot agent's psychological attitudes to affect the robot's outside, physical, nonsymbolic environment, and *vice versa.* It is only the hyper-logicist who jettisons such transducers — as we shall see below.

At any rate, the Fodorian argument seems to us to be unsuccessful, for the simple reason that, pointed out to a large degree by (Garson 1990), recent advances on the connectionist front (c.f. (Servan-Schreiber, Clceremans et al. 1988), (Elman 1989), (Kaplan, Weaver el al. 1990)) have resulted in systems that model the abilities thought by Fodor and company to be symbolic in nature. This development would seem to be thoroughly unsurprising, because it would seem to be just what the formal results ensure. If one puts no artificial limit on type or complexity of a neural net, then you quickly have the |i-recursive functions available, and therefore you have Turing computability. And the converse holds also: for every Turing machine you can be sure there is a neural net that matches it The mathematics of the situation, specifically the ultimate equivalence of neural nets and Turing machines, would seem to doom forever the Fodorian tack. We should before long have connectionist systems on the scene that are very good at handling those aspects of human language Fodorians hold to be the special province of logicist approaches. This is because, in a genuine sense, neural nets are just Turing machines.

The point here is based on the locution 'automaton *x* is just automaton y,' which may be used to encapsulate such facts as that cellular automata arc just it-tape Turing machines, or that Register machines (for an intro sec (Ebbinghaus, Flum et al. 1984)) are equivalent to Turing machines. The idea here is that, ultimately, from the mathematical point of view, *x* and *y,* in the locution being considered, are the same creature: you could in principle specify both by the exact same set theoretic definition, starting from and never leaving the machinery of, say, ZFC.

This may look like an obvious point, but recent work on the C-L debate flies in its face, since even those who explicitly consider the point seem to go to considerable lengths to dodge it. Here, for example, is what (Harnad, 1990) has recently said on the matter:

> There is some misunderstanding of [the "fact" that neural nets fail to meet certain necessary conditions for a symbol system] because it is often conflated with a mere implementational issue: Connectionist networks can be simulated using symbol systems, and symbol systems can be implemented using a connectionist architecture, but that is independent of the question of what each can do qua symbol system or connectionist network, respectively. By way of analogy, silicon can be used to build a computer, and a computer can simulate the properties of silicon, but the functional properties of silicon are not those of computation, and the functional properties of computation are not those of silicon.

A proper analysis of this rather cryptic quote would require clarification of nothing less than the notions of functional properties, analogical arguments, and simulations. Such clarification is an impossibly tall order, certainly given our space limitations. And yet it is easy to motivate such clarification: What are the functional properties of pencils? Of pencils if some race who shun writing find them and use them to spin frisbees upon? Fortunately, such questions need not detain us. We think it can rather easily be seen that Harnad cannot be construed as here threatening the locution under scrutiny: Suppose that we have an operator '$[\![\;]\!]$' which when applied to a standard, "user friendly" specification of an automaton or neural net, yields an account expressible exclusively in ZFC. Then we are relying on the proposition that

> For every neural net N, there is a Turing machine *M* such that $[\![M]\!] = [\![N]\!]$.

We take it, furthermore, that for every *x* and y, if *x* = y, then *x* and y have precisely the same functional properties. New York's tallest building has the same functional properties, no matter what such properties amount to, as the World Trade Center. This follows from Leibniz' Law.

So it looks like the Fodorian case, indeed any case which involves a claim about the "qualitative" differences between neural nets and conventional automata, is misguided. But we can nonetheless get appreciably clearer about what strong logicism is. Here, in fact, is a stab at doing so: strong logicism, minimally, seems to us to include the following two propositions:

(PER_{TUR} Persons are Turing machines.

(SYM) If a robotic person *S* is to be eventually produced by logicist Alniks, then *S* must be such that some of the propositions ϕ_0, ϕ_1, ... which are objects of S's occurrent deliberations (and hopes, fears, etc.) are represented by formulas «ϕ_0», «ϕ_1», ... of some symbol system $\mathcal{L}^{\mathcal{T}}$, where they can be processed according to the reasoning mechanism that is part of $\mathcal{L}^{\mathcal{T}}$.

The identification of strong logicism with, minimally, these two theses is not something pulled out of thin air. (PER$_{TUR}$) reflects not only the strong logicist's view that she is ultimately aiming at the creation of a true agent, but also the view that this agent would be, at bottom, a classical, or *symbolic,* automaton. One can say that a classical automaton is a symbolic automata, because every classical automaton is identical to an axiom system within some symbol system. This fact is what allows the proof of the undecidability of first-order logic to capitalize on the halting problem. (SYM), on the other hand, reflects the logicist tenet (cf. (Fodor 1980, 1985), (Fodor & Pylyshyn 1988)) that symbol strings of a symbol system (in this case the imaginary $\mathcal{L}^{\mathcal{T}}$) capture what mental phenomena such as thoughts and beliefs *arc.* The view here is that there is a level of thought, the "mental" level, with ruleful regularities that are independent of their specific physical realizations.

We see these propositions as *necessary, not* sufficient, for strong logicism. One of the missing propositions might be a general thesis stating what "theory of mind" is operative in strong logicism. One candidate found in the literature for a theory of mind closely allied with traditional logicist AI is called AI-Functionalism' by (Rey 1986). Viewed in the "flow chart" terms of (Dennett 1978), AI-Functionalism says, intuitively, that if you find a flow chart match between human brains and silicon-based Martian brains, then you can be assured that the human person and the Martian enjoy the same mentality.

Our characterization of strong logicism provides a springboard for describing hyper-logicism, and to it we now turn.

3 Hyper-Logicism

What, given the foregoing, is hyper-logicism? Since, as far as we know, no one has explicitly articulated or championed this view in the literature, this question cannot be answered by simply thumbing through the appropriate paper or book. We nonetheless think we have some ideas about what hyper-logicism is about. Specifically, it is for us the view which embraces at least the following three propositions:

(PER$_{TUR}$) Persons are Turing machines.

(SYM!) AI can produce a robotic person *S* such that *all* of the propositions ϕ_0, ϕ_1, ... which are objects of S's deliberations (and hopes, fears, etc.) are represented by formulas «ϕ_0», «ϕ_1», ... of some symbol system $\mathcal{L}^{\mathcal{T}}$, where they can be processed according to the reasoning mechanism of $\mathcal{L}^{\mathcal{T}}$.

CNO-SUfi) AI can produce a robotic person *none* of whose mental processing involves *subsymbolic* encodings not representable in $\mathcal{L}^{\mathcal{T}}$.

In the following three sections, as promised, we will defend these theses.

4 Agents as Classical Automata

To argue directly for (PER$_{TUR}$) would mean, among other things, confronting head-on questions about the ability of human persons to exceed the algorithmic. Indeed, a cogent case for (PER$_{TUR}$) would of necessity be a long, sustained essay in the philosophy of mind. Fortunately, we can dodge the burden: there is no reason why we must take on the onus of proving (PER$_{TUR}$). This is so because one of the assumptions behind the C-L debate is that agents *just are* to be understood in computational terms. The issue, in the present context, is *which* terms. After all, the strong logicist is also saddled with having to defend (PER$_{TUR}$), and the strong conncctionist may face the challenge of having to show that persons are neural nets.

Let us make an enabling assumption: that despite our above objections to the Fodorian argument for strong logicism, there *is* a substantive difference between on the one hand a thesis claiming that persons arc Turing machines, and on the other that they arc, say, neural nets. Then we can ask a question which gives rise naturally to the onus we *should* be under in the present section: How might differences between Turing machines and neural nets end up supporting conncctionists? Well, perhaps the connectionist would try to capitalize on the fact that neural nets are analog devices. The conncctionist might say that though neural nets and cellular automata and k-tapc Turing machines are one and the same when considered through the lens of operators like '$[\![\ ]\!]$,' when these automata arc genuine *physical* entities in the *physical* world they arc quite different; and their differences could be, from the standpoint of generating mentality, significant.

Yet this is a remarkable position. In order to see this, consider the following situation. Suppose that we have a *physical* neural net, call it 'N^*,' that computes a set of functions Γ; and suppose that N^* has been built out of stuff available in the physical world to connectionists. Suppose that this neural net is very complex, closer by far to real human brains than to standard textbook diagrams of multi-layer nets. And now suppose that, using N^* as a sort of blueprint, we build a Turing machine M^* that computes all of Γ. If we had the time, we could specify how M^* is to be built from N^*. For example, suppose N^* is a 50 layer neural net, and that input "neurons" are 1000 in number; then we might want to build M^* as a 50-tape machine, with 1000 squares of the first tape used to hold the input that goes into *N**. And so on.

Now. Here is the crucial question: Is it plausible to hold that while the immaterial set-theoretic versions of *N** and M* amount to the same thing, i.e. that [[*N*]]* = [[M]], the *physical* versions don't? That the net and the Turing machine here don't give rise to the same mental states (if in fact there *are* any in the picture)? It may seem at first glance that there is no rationale supporting an affirmative answer to these questions. But this would be to move too quickly; it would be to ignore the importance the connectionist is placing on the physical. For this view might very well include

(ANA.) A true analog, neural net can compute things which no Turing machine can.

and thereby imply a rejection (= the falsity) of

(CTT*) Whatever can be accomplished by a computing machine of any sort, can be accomplished by a suitably programmed Turing machine.

And rejecting (CTT*) allows one to hold that while our physical net *N** from above computes r, there is no *physical* Turing machine that can compute T; and if there can be no such Turing machine, then our little thought-experiment involving *N** and A/* is all for naught. So in this response the connectionist affirms AI-Functionalism, but also embraces a positive thesis about what sort of *physical stuff* is of paramount importance — stuff that can't be matched, functionally speaking, by any Turing machine.

Have we arrived, then, at a solid rationale for refusing to accept (PER_{TUR})?

Well, if nothing else, this version appears to reflect the current situation. As of 1991, nearly all neural networks arc implemented on general purpose parallel computers — computers whose power is specified, mathematically, by cellular automata. Cellular automata, as we have noted, when viewed from the perspective of the foundations of mathematics, are exactly equal in power to Turing machines. Hence as of 1990 neural computers can be viewed as Turing machines, and it follows that today whatever can be done by a neural net can be done by an ordinary Turing machine. But in light of this result our connectionist calmly proclaims that hardware *is* all-important in reaching AI's ultimate goals, not solely in the sense of moving toward "brainlike" architectures; but hardware is all-important for the simple reason that wc don't *really* have a neural net as long as we are forced to implement it on a programmable, general purpose machine. We will have a *true* neural net, the strong connectionist continues, when and *only* when we implement a neural net which is isomorphic to that underlying the human brain on a *true analog machine.*

What are we to make of the position that the connectionist is now occupying? We are inclined to view the situation here as calling for a big application of *modus tollens.* That is, since we affirm both (CTT*), and since the present version of connectionism entails the *negation* of this proposition, we think that this version of connectionism is simply false.

Now we haven't the time to consider arguments for and against (CTT*). It is at the very least inductively confirmed by the fact that researchers have never found a computing machine, whether analog or not, that is qualitatively superior to a Turing machine. And while in principle a counter-example to (CTT*) is possible, no one takes this prospect seriously. There is also the fact, only recently noted by (Mendelson 1990), that the Church-Turing Thesis and its relatives may, in a sense in use in mathematics, be provable.

But might there be other formidable reasons for an AInik to resist (PER_{TUR})? Perhaps. One might say that neural nets, while "qualitatively" equivalent to Turing machines via '[ ],' are nonetheless superior to Turing machines "quantitatively." This amounts to saying that neural nets are, in terms of complexity, superior to Turing machines, when the problems in question are those crucially involved in thought at the heart of "agenthood." This is a vague stance (which nets are competing with which Turing machines? ...), but one worth taking seriously in the present context. Can it succeed against the hyper-logicist? We don't think so. Here's why.

There now seems to be reason to think that conventional automata, suitably realized, could be phenomenally powerful complexity-wise. We are referring to aspects of what are being called "quantum computers" (cf. (Lockwood 1989), (Penrose 1989)). While the physics behind these devices involves the difficult and, to many, puzzling field of quantum mechanics, and while most of those who discuss quantum computers appear to affirm the highly speculative thesis that the brain is a quantum computer, it does seem that the purely mathematical specification of a quantum computer is unexceptionable. Whatever else one might say about quantum mechanics, it is surely the case that the mathematical techniques employed within it arc above reproach. A quantum computer, in exclusively formal terms, severed from picturesque claims about how the brain might be one, does appear to be a genuine generalization of standard Turing machines. It appears that, for certain problems P_i that cannect be encoded as functions, Turing machines cannot solve P_i in polynomial time proper, whereas quantum computers can (cf. (Lockwood 1989)). The basic idea behind quantum computers — let us call them Q-machincs — is that they are as individuals "groups" of conventional Turing machines whose internal states and tape states can be superpositions of these Turing machines.

Wc are not claiming here that g-roachines can be built; nor are we claiming that the brain is a Q-mahinc. Our point is only that, *for all we know at the present time about what robot agents must be like,* it is permissible to interpret (PER_{TUR}) as referring not to Turing machines, and not to Q-midlines, but to what might be called "generic classical machines," where these automata could be Turing machines, Q-machines, or some other exotic variation on the classical theme. This moves seems to us to take the wind out of the connectionist's sails, since there would seem to be little *a priori* reason to be sanguine about neural nets *over and above* classical automata for reasons pertaining to complexity.

At any rate, what is distinctive about hyper-logicism are (SYML) and (NO-SUB), to which we now turn.

5 The Lesson of the Parallel Postulate

The central claim of this section is that consideration of the lesson physics has taught us by way of alternative qeometries provides significant reason for embracing (SYML!). Perhaps the quickest way to sec this is to begin by considering arguments that would typically be brought *against* this thesis. Such arguments (Kaplan Weaver et al. 1990) almost invariably appeal to the claim, allegedly established by certain psychological experiments (cf. (Nisbett and Ross 1980), (Kahneman & Tversky 1973) (Margolis 1979)), that human persons hardly ever use logic to reason, and that when they are given problems which require the use of logic, subjects use it egregiously. We have have doubts about the validity of these experiments *as evidence against the thesis that human persons competently employ logic.* Nonetheless we are willing for the sake of argument to concede that these experiments have the implications many connectionists claim they have. What we would like to suggest, however, is that these connectionist claims reflect the same sort of parochial attitude that used to exist about non-Euclidean geometries. Let us explain.

We begin with a compressed timeline from Euclidean Geometry to Physical Geometry: There was uneasiness about the fifth, or parallel, postulate (PP) for at least 2000 years, during which time there were sporadic attempts to deduce it from the other four postulates. In 1733, Saccheri constructs an indirect proof by denying PP, which yields two routes, "no parallels," and "many parallels." The first case is inconsistent with the other four postulates, but the second is not. Saccheri discovers a non-Euclidean geometry but convinces himself that such a thing is absurd and contradictory. Around 1800 Gauss demonstrates to his own satisfaction the possibility of many-parallels non-Euclidean geometry, but doesn't publish the result. In 1820 Bolyai and Lobachevski independently publish a many-parallels geometry. In 1850 Riemann works out a no-parallels geometry, as well as a generalized geometry which has Euclidean (one parallel), Bolyai-Labachevski (many parallels) and Riemannian (no parallels) as special cases. In 1905 Einstein's Special Theory of Relativity connects measurements of space, time, and motion with light signals and revolutionizes the concepts of space and time in physics. In 1915 Einstein's General Theory of Relativity gives a theory of gravitation which equates gravity with the structure of space-time. They theory makes use of generalized four-dimensional Riemannian geometry and shows how the local Euclidean character of space is a consequence of scale. The overall structure of the universe is non-Euclidean, a finite but unbounded four-dimensional manifold.

Wc are claiming that there are lessons here which may (partially, anyway) adjudicate the C-L debate in favor of hyper-logicism. In order to see this, suppose first that "mindspace" is thought of in terms of physical space. Ordinary interaction with physical space, given that this interaction is neither with the very small nor with the very large, suggests a classical Newtonian conception — one that is useful for sending rockets to the moon, but one which is not universally true. Perhaps what we see of mindspace is similarly restricted. Perhaps there are ways of thinking that differ substantially from our ways. Perhaps, though we can't handle 17 nested quantifiers, there are cognizcrs who can. And so on.

There seem to us to be four general lessons which can be drawn from the physics story:

(L1) Be cautious about extrapolations from local situations.

(L2) What we can or cannot conceive, imagine, or visualize at any given time is not a reliable indication of what is possible or actual.

(L3) An easy separation of "conceptual" and "empirical" can be misleading. What appears to be clearly a conceptual or logical issue might have hidden empirical aspects and what seems to be a straightforward empirical problem might have subtle conceptual connections.

(L4) It takes time to work things out.

How might these lessons be applied to the C-L debate in such a way that hyper-logicism gains credence? Let's take the lessons in turn, starting with (LI).

(LI) is based on the fact that physics has shown us that we cannot move from our sense of local simultaneity to absolute simultaneity and the notion of a "universal now," nor can we extend a local Euclidean framework to galactic distances. The moral for AI would seem to be that we should be wary about generalizing from our local conception of mentality to absolute mentality. To be more specific, we should leave open the possibility not only that there are alternative modes of cognition which not only make crucial use of symbol systems, but which use symbol systems exclusively when it comes to "propositional attitudes." Such a view does not allow one to generalize from the typical, "local," "pro-connectionist" psychological experiment (in which, say, humans, given tasks thought naturally to require the use of logic, *dont* use logic) to the proposition that all agents are similarly inept at using logic.

(L2)'s import, in the present context, seems clear: while wc perhaps cannot conceive of what it would be like, when cognizing at the level of occurrent propositional attitudes, to exclusively employ a symbol system for such cognition, this fact (if it is a fact) should not be taken as a reliable indication that a being whose cognition is couched exclusively in some symbol system is impossible.[1]

(L3) has perhaps the most interesting and specific implications for hyper-logicism and the C-L debate. The point here would seem to be that just as non-standard positions on the parallel postulate, itself conceptual and formal in nature, yielded genuine empirical possibilities, non-standard positions on formal issues related to symbol systems may yield genuine possibilities for hyper-logicist cognition. When logicism is criticized, it is almost invariably true that what is being criticized is standard first-order logic. This is like attacking physics by pointing out that the Euclidean scheme cannot accommodate empircal puzzles now explained by Relativity Theory. It is now well-known that if first-order logic is abandoned, myriad alternative possibilities open up. The first batch of such possibilities derives from second order, indeed /i-order, extensional logics. Even the simplest of second-order logics produce a significant increase in expressive power (the Peano axiom system can be formalized in second-order logic). And there are other possibilities: First-order logic is monotonic; dropping this restriction appears to hold promise for modelling cognition thought by many to be beyond the reach of a symbol system (for a competent distillation sec (Nilsson & Genersereth 1988)). As mentioned above, there is intensional logic. And then of course there is an infinite number of as-yet undiscovered possibilities for symbol systems.

The lesson of (L3) can perhaps be put most forcefully in the context of a concession sometimes made by connectionists. The concession here, made e.g. by (Hamad 1990), is that some mental activity, say for example the conscious, sustained "in the head" manipulation of a symbol system and its components (something often done by logicians), is *by its very nature* symbolic. What we want to suppose, in keeping with (L3), is that there may be, for all we know as residents of local mindspace, agents that are not only inhumanly good at this kind of symbolic activity, but agents whose cognition is couched *exclusively* in such symbolic reasoning.

(L4), finally, has a very simple message. The timeline for physics suggests to us that dissatisfaction with the logicist approach to AI is remarkably impatient. Logicians and mathematicians working in AI may need to toil for centuries before their work can be used to undergird implementation that leads to the creation of a genuine robotic agent. If anything, mindspace would seem to be a more difficult domain than physical space. It would thus come as no surprise if the timeline for logicist AI stretched on well into the future.

The physics story, then, appears to us to lend credence to (SYM). What about the final member of the hyper-logicist triad? We turn to a defense of it now.

6 Brains-in-a-Vat Thought Experiment

The (NO-SUB) part of the hyper-logicist triad is the view that AI should proceed in the hope of building robot agents resembling those in (Putnam 1981)'s brain-in-a-vat thought-experiment. These agents would have no sensors and no way of changing the external world, and would have no need (so the story would go) of subsymbolic processing so well-suited (as the connectionists have shown) to handling the relation between an agent and the world through which it navigates. On the other hand, hyper-logicism, if tenable, leaves room for cognition that, as Putnam and others have shown, can be in many ways as rich as our own.

At the core of our defense of (NO-SUB), then, is a thought-experiment — which runs as follows. The year: 2047. Sue, at birth, is pared down to her nascent brain and then placed in a vat instead of her mother's arms. This vat, supervised by an ingenious neuro-Alnik, is actually an extraordinary device capable of providing, *via* a massive webwork of electrodes, all the input that normally enters a brain through standard sensory pathways. Sue, "growing up," gets all the input we get by way of a rather sinister short-cut. But what she "sees" is as vivid as what we see, and what she "hears" is an clear as what we hear; and so on.

One might argue, as (Putnam 1981) does, that Sue could not be a person in the full sense; that *we* could not be, as some philosophical skeptics have claimed, brains in vats. We are willing to agree that such an argument could be made out, and that it would be sound. Our point, however, concerns (NO-SUB): wc are claiming that the thought-experiment shows that

(1) O There exists a robotic agent *none* of whose mental processing involves *subsymbolic*

[1] Our point here can perhaps be bolstered by a distinction between :nternal" and "external" visualization, due to the 19th century philosopher-scientist H. von Helmholz. We can only sketch the argument here: External visualization makes use of one dimension of space to visualize arrangements of other dimensions. Thus, we have no difficulty in visualizing two-dimensional manifolds as embedded in a three-dimensional space and we suppose that there are no problems in visualizing a Euclidean space of three dimensions. But surely there are different kinds of visualization; we cannot avail ourselves of a fourth dimension to visualize three-dimensional situations embedded in a four-dimensional manifold. And here is where *internal* visualization enters: to visualize space internally is to imagine the kinds of experience one would have if she were living in such a space.

Thus, with respect to visualization, both Euclidean and non-Euclidean spaces require internal visualization. This seems easy for the Euclidean case because we live in this space and the non-Euclidean cases arc correspondingly difficult to visualize because of our lack of practice. But, in principle, we can visualize the kinds of experience we would have if we lived in such a space. And, in fact, modern cosmology is a sort of aid to the development of a non-Euclidean imagination! The "privileged" epistemological status of Euclidean geometry turns out upon analysis to depend on psychological habituation — we are more familiar with Euclidean space and seem to slide easily from external to internal visualization when, in fact, there are different kinds of "visualization." Euclidean space has no privileged logical or epistemological status but simply the privilege of habit. With respect to the internal visualization of non-Euclidean space, wc only lack practice.

encodings (traditionally required for agent-environment interaction) not represcntable in L^{J}

This robot agent of (1) needn't be just like *human* agents.

Many questions remain; we have space enough to briefly consider only two. The first is this: Is it really true that

(2) $(1) \rightarrow (\text{NO-SUB}),$

given that (NO-SUB) says that logicist AI can actually *produce* the sort of robot agent in question? We think the answer to this question is "Yes," because Sue-like robot agents would seem not only logically possible, but *physically* possible, and not only physically *possible,* but *buildable.* Let (T) be the proposition produced by modifying (1) in this way (changing ◇ to ⟐, the latter 'physically possibly'). It seems us, then, that

(2') $(1') \rightarrow (\text{NO-SUB}),$

And so if we are right that Sue and her more robotic relatives confirm (1*), we will have the third and final member of the hyper-logicist triad by *modus ponens.*

Here is the second question: What would be the *point* of trying to build a robot agent not connected to the external world? Some philosophers may balk at this question, because philosophers seem to specialize in precisely the kind of thinking that has very little to do with well-defined neural structures, and everything to do with the kind of thing that one can do while in a sensory deprivation tank, or do limbless, paralyzed, and all alone. But there is a much less pedantic and sensible way to address the question: It seems plausible to think that there arc problems in the world which might be solvable by a robot agent engaged in symbolic thought in the absence of any "hook ups" to the external physical world (save for a keyboard or some such device to bring symbol strings directly in and out). We have in mind complex, symbolic macroeconomic and microeconomic problems the solving of which would presumably have great utility for human life, but the reader can no doubt think of other examples.

References

Churchland, P. M. and P. S. Churchland. (1990). "Could a Machine Think?" *Scientific American.* (January): 32-37.

Dennett, D. (1978). *Brainstorms.* Bradford Books.

Ebbinghaus, H. D., J. Rum and W. Thomas. (1984). *Mathematical Logic.* New York, Springer Verlag.

Elman, J. (1989). *Representation and Structure in Connectionist Models.*

Fodor, J. (1980). "Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology." *Behavioral and Brain Sciences* 3: 63-109.

Fodor, J. (1985). "Precis of *The Modularity of Mind" Behavioral and Brain Sciences.* 8: 1-42.

Fodor, J. and Z. Pylyshyn. (1988). "Connectionism and Cognitive Architecture: a Critical Analysis." *Cognition.* 28: 3-71.

Garson, J. (1990). "Cognition without Classical Architecture." (unpublished manuscript).

Harnad, S. (1990). "The Symbol Grounding Problem." *Physica D* 42 335-346

Kahneman, D. and Tversky, A. (1973) "On the Psychology of Prediction." *Psychological Review,* 80, 237-251.

Kaplan, S., M. Weaver and R. French. (1990). "Active Symbols and Internal Models: Towards a Cognitive Connectionism." (unpublished).

Lockwood, M. (1989). *Mind, Brain, and the Quantum.* B. Blackwell.

Margolis, H. (1979). *Patterns, Thinking, and Cognition.* Chicago, IL: University of Chicago Press.

Mendelson, E. (1990). "Second Thoughts about Church's Thesis and Mathematical Proofs." *Journal of Philosophy.* May.

Nilsson, N. and M. Genesereth. (1988). *Logical Foundations of Artificial Intelligence.* Morgan Kaufman.

Nisbett, R. and Ross, L. (1980). *Human Inference: Strategies and Shortcomings of Social Judgement.* Englewood Cliffs, NJ: Prentice-Hall.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford University Press.

Putnam, H. (1981). "Brains in a Vat" *Reason, Truth and History.* Cambridge University Press.

Rey, G. (1986). "What's Really Going on in Searles 'Chinese Room'." *Philosophical Studies.* 50: 169-185.

Rumelhart, D. and J. McClelland. (1986). *Parallel Distributed Processing.* MIT Press.

Servan-Schreiber, D., A. Cleeremans and J. McClelland. (1988). *Encoding Semantical Structure in Simple Recurrent Nets.*

Smolensky, P. (1988). "On the Proper Treatment of Connectionism." *Behavioral and Brain Sciences.* 11: 1-74.

Waltz, D. (1988). "The Prospects for Building Truly Intelligent Machines." *The Artificial Intelligence Debate.* MIT Press.