Delft University of Technology

Explainable Artificial Intelligence for human-AI collaboration

van der Waa, J.S.

DOI
10.4233/uuid:855bd55f-d297-4fff-9248-9cde0ad2b976

Publication date
2022

Document Version
Final published version

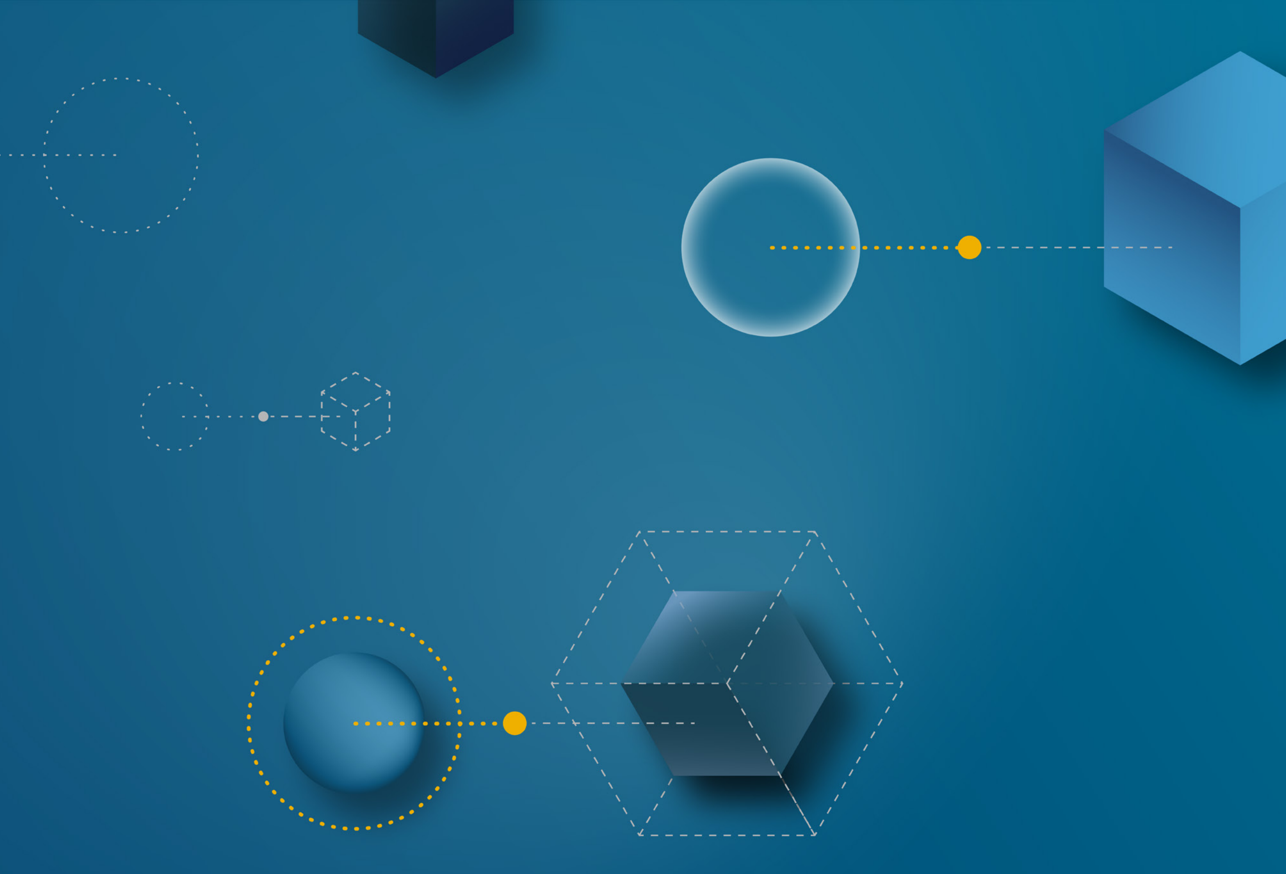
Citation (APA)
van der Waa, J. S. (2022). *Explainable Artificial Intelligence for human-AI collaboration*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:855bd55f-d297-4fff-9248-9cde0ad2b976

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.



EXPLAINABLE ARTIFICIAL INTELLIGENCE

FOR HUMAN-AI COLLABORATION

Jasper van der Waa

EXPLAINABLE ARTIFICIAL INTELLIGENCE

FOR HUMAN-AI COLLABORATION

- - - - - - - - - - - - - - - -

Explainable Artificial Intelligence for human–AI collaboration

Dissertation

for the purpose of obtaining the degree of doctor
at Delft University of Technology
by the authority of the Rector Magnificus, prof. dr. ir. T.H.J.J. van der Hagen,
chair of the Board for Doctorates
to be defended publicly on
Wednesday 19 October 2022 at 12:30 o'clock

by

Jasper VAN DER WAA

Master of Science in Artificial Intelligence,
Radboud University, Nijmegen, the Netherlands,
born in Heerlen, the Netherlands.

This dissertation has been approved by the promotors.

Composition of the doctoral committee:

Rector Magnificus	Chairperson
Prof. dr. M.A. Neerincx	Delft University of Technology, promotor
Prof. dr. C.M. Jonker	Delft University of Technology, promotor

Independent members:

Prof. dr. F.M. Brazier	Delft University of Technology
Prof. dr. ir. J.F.M. Masthoff	Utrecht University
Prof. dr. A. Nowé	Free University of Brussels
Prof. dr. S.A. Raaijmakers	Leiden University
Prof. dr. ir. A. Bozzon	Delft University of Technology, reserve member

Other member:

Dr. J. van Diggelen	Netherlands Organisation for Applied Scientific Research

Keywords Artificial Intelligence, Explainable AI, Machine Learning, Human-AI Collaboration
Printed by Ridderprint
Cover design Persoonlijk Proefschrift
Layout design Youetta Kunneman

Copyright © 2022 by J.S. van der Waa

ISBN 978-94-6458-619-0
An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

TABLE OF CONTENTS

Part I - Explanations for understanding

Part II - Explanations to act upon

Part III - Explanations in human-AI collaboration

Part IV - Conclusions and discussion

SUMMARY

As a society, we have come to notice the influence and impact Artificially Intelligent (AI) agents have on the way we live our lives. For these AI agents to support us both effectively and responsibly, we require an understanding on how they make decisions and what the consequences are of these decisions. The research field of Explainable Artificial Intelligence (XAI) aims to develop AI agents that can explain its own functioning to provide this understanding. In this thesis we defined, developed, and evaluated a core set of explanations an AI agent can provide to support their collaboration with humans.

In Part I we studied the effects of explanations that convey why one decision was made instead of another, i.e., the contrastive explanation class. Two forms of this class were evaluated (Chapter 2), providing either rule-based or example-based content. The rule-based form improved a human's understanding the most. Both explanations caused participants to feel they understood the AI agent, although this did not correlate with their actual understanding. Furthermore, with a self-explaining agent the participants proved to be more persuaded to follow the agent's advice even when incorrect, particularly when the explanations were provided in an example-based form. A method to generate rule-based contrastive explanations was developed for AI agents that offer decision support and proven to be efficient, accurate and agnostic from the AI agent's functioning (Chapter 3). Based on our findings of a pilot study, a second method was presented and defined for AI agents that plan behaviour over time, as often used in autonomous systems (Chapter 4). These findings indicated that humans desire contrastive explanations from such planning AI agents to report what consequences that agent expects when performing its plan. The presented and defined methods takes this into account by allowing humans to question the plans of these AI agents and receive an answer addressing the agent's expected consequences conform human interpretable terms instead of numerical values (i.e., that by turning right the agent expects to fall off a cliff, instead of explaining that turning right reduces the expected utility significantly).

In Part II we defined two novel explanation classes: confidence explanations and actionable explanations. Confidence explanations convey the likelihood of an AI agent's decision to prove correct, compute this in an interpretable way and explain it using past examples of performance (Chapter 5). We proposed an agnostic approach to generate such explanations using case-based reasoning. Evaluations showed this approach to be accurate and predictable, even under simulated updates of the AI agent and concept drift over time. Two studies showed that both laypeople and domain experts preferred our case-based reasoning approach for confidence explanations over state-of-the-art alternat-

ives. Actionable explanations aim to support a human's ability to contest and alter an AI agent's decision when that human is subjected to the AI agent's decision (Chapter 6). We formally defined six properties that make an explanation actionable to enable univocal comparisons and argumentations on explanation theories that contribute to contesting AI agents' decisions. A literature review was performed to provide a research agenda for the development and testing of methods to generate actionable explanations.

Finally, in Part III we recognize that explanations serve the collaboration between humans and AI agents and their application needs to be designed within the context of that collaboration. We extended an existing design method for such collaborations with the notion of explanations and presented several designs (Chapter 7). Each design varied in the degree of autonomy provided to the AI agent in morally sensitive tasks and discussed the role of explanations within such tasks. Several of these designs were then evaluated in the healthcare domain with first responders (Chapter 8). Results showed that the participants valued the explanations but also found them tedious when experiencing time pressure. Furthermore, they felt less responsible for the AI agent when it became more autonomous which reduced their motivation to review the explanations. This illustrates the complexity of designing an explainable AI agent that integrates various explanations to support a human–AI collaboration.

To summarize, the above findings show that explanations from an AI agent have the potential to improve the collaboration between human and AI agents since explanations can bring about various beneficial effects. Not all these effects are positive, however. Explanations can also induce negative effects detrimental to the collaboration, which is dependent on context. For instance, a more persuasive advice due to an explanation might be detrimental in a use case where it prohibits a desirable critical human stance but beneficial when it remedies unwarranted under trust. The performed studies showed that explanations induce effects whose value is use-case dependent. Similar future studies measuring the variety of effects explanations bring about will provide a solid foundation for design choices on explainable AI agents. Such design choices could be structured and made accessible with the use of design patterns that describe which explanations have what effects in what kind of use cases given a particular kind of human–AI collaboration. Aside from these insights, we illustrated the value of a more formalized approach towards explanations that are actionable instead of only having an epistemic value. Through distinct properties and levels, we could provide a research agenda towards explanations with the profound practical value of enabling human autonomy when interacting or dealing with AI agents. Finally, we showed that the development of explanation generating methods that are independent of the applied machine learning technique can be effective. This is especially the case when explanations do not need to disclose every detail of an AI agent.

We conclude with advice for future research in XAI. First, our advice is to spend more effort on the evaluation of the explanations as well as the theoretic foundation on how such explanations are generated. There is a need for more rigorous evaluations grounded in realistic applications of AI agents based on explicit theoretical models describing both positive and negative effects of explanations. In addition, with more attention to the theoretical and mathematical foundation on how explanations should be generated, we can work towards methods for which we are equivocally able to determine when the generated explanation is sufficiently correct given the intrinsically complex AI agent it explains. With an increased effort in both, we should be able to provide industries and governments with the knowledge needed to apply explanations effectively and responsibly and to formulate best practices and regulation. Our second advice is to provide more attention to the role and embedding of explanations in the human–AI collaboration. This will open more purposes for explanations instead of solely for creating trustworthy AI agents. Explanation purposes such as supporting long-term collaborations, aiding in knowledge discovery, educating humans in a domain. This advice comes down to a more profound focus on human-centred research to the explanations an AI agent should provide.

SAMENVATTING

Als samenleving zijn we de invloed en impact gaan merken die kunstmatig intelligente (AI) agenten hebben op onze manier van leven. We hebben inzicht nodig in hoe AI-agenten hun beslissingen maken en de consequenties hiervan, willen we dat deze AI-agenten ons effectief en verantwoord ondersteunen. Het onderzoeksveld Explainable Artificial Intelligence (XAI) heeft tot doel AI-agenten te ontwikkelen die hun eigen werking kunnen uitleggen en hiermee dit benodigde inzicht te verschaffen. Dit proefschrift adresseert verschillende soorten uitleg die een AI-agent kan geven om zo de samenwerking met mensen te ondersteunen. Deze soorten uitleg worden gedefinieerd, ontwikkeld en geëvalueerd.

In Deel I hebben we de effecten bestudeerd van een uitleg waarom een AI-agent de ene beslissing is nam in plaats van een andere, de zogenoemde contrastieve uitlegklasse. Twee vormen van deze klasse werden geëvalueerd (Hoofdstuk 2). Een vorm gebaseerd op beslisregels en een vorm gebaseerd op gedragsvoorbeelden. De op regels gebaseerde vorm bleek het begrip het meest te bevorderen in een uitgevoerd experiment. Beide verklaringen zorgden ervoor dat de deelnemers het gevoel hadden dat ze de AI-agent begrepen, hoewel dit niet correleerde met hun daadwerkelijk verkregen begrip. Bovendien lieten deelnemers zich vaker overtuigen om het gegeven advies op te volgen als de AI-agent dit advies kon uitleggen, zelfs als dit advies onjuist was. Dit gebeurde met name wanneer de uitleg gebaseerd werd op voorbeelden. We ontwikkelde een methode om op regels gebaseerde contrastieve uitleg te genereren voor AI-agenten die mensen adviseren over een te nemen besluit. We toonde aan dat deze methode efficiënt, nauwkeurig en onafhankelijk fungeert van het functioneren van de AI-agent (Hoofdstuk 3). Op basis van onze bevindingen van een pilotstudie werd een tweede methode gepresenteerd en gedefinieerd voor AI-agenten die plannen maken over de tijd, zoals autonome systemen (Hoofdstuk 4). De bevindingen van deze pilotstudie gaven aan dat mensen van plannende AI-agenten een contrastieve verklaringen willen die verklaart welke gevolgen de AI-agent verwacht bij het uitvoeren van het plan. Een methode is gepresenteerd die dergelijke uitleg kan genereren door mensen in staat te stellen het gegenereerde plan in twijfel te trekken en een alternatief voor te stellen. Hierop genereert de voorgestelde methode een uitleg waarom de AI-agent niet voor dit alternatieve plan koos in termen van het verschil in gevolgen. Deze gevolgen worden gegeven in door mensen te interpreteren termen in plaats van numerieke waarden (bijvoorbeeld, dat als de agent naar rechts zou gaan deze van een klif zal vallen, in plaats van uit te leggen dat rechts afslaan het verwachte nut aanzienlijk vermindert).

In Deel II definieerde we twee nieuwe klassen van uitleg: uitleg over zekerheid en actiegerichte uitleg. Een uitleg over zekerheid van de AI-agent communiceert en verklaart wat de kans is dat een beslissing of advies van de AI-agent correct zal zijn. De zekerheid moet op een interpreteerbare wijze worden berekend en uitgelegd worden aan de hand van besluiten uit het verleden (Hoofdstuk 5). We ontwikkelde een methode die onafhankelijk werkt van de AI-agent die deze berekening en uitleg kan geven door te redeneren over verleden gedragsvoorbeelden. Evaluaties toonden aan dat deze methode zowel nauwkeurig is als voorspelbaar, ook als we simuleerde dat de AI-agent plots van gedrag verandert door een update of als er concept-drift optreedt over de tijd. Twee studies toonden daarnaast aan dat zowel leken als domeinexperts de voorkeur geven aan de door ons gedefinieerde uitleg van zekerheid dan andere alternatieven. Actiegerichte uitleg zijn bedoeld om mensen te ondersteunen in hun vermogen om de beslissing van een AI-agent te betwisten en te wijzigen. In het bijzonder wanneer de mens onderworpen wordt aan deze beslissing (Hoofdstuk 6). Zes eigenschappen van een dergelijke uitleg zijn gedefinieerd en geformaliseerd. Een uitleg die voldoet aan alle zes, maakt deze actiegericht. Deze zijn geformaliseerd om eenduidige vergelijking en argumentaties mogelijk te maken in het onderzoeksveld over de soorten uitleg die mensen in staat stellen om besluiten van AI-agent te betwisten. Een onderzoek agenda is opgesteld op basis van een literatuuronderzoek en deze zes eigenschappen om de ontwikkeling van actiegerichte uitleg te ondersteunen.

Ten slotte erkennen we in Deel III dat verklaringen de samenwerking tussen mensen en AI-agenten moeten dienen en dat hun toepassing binnen de context van die samenwerking zal moeten worden ontworpen. Een bestaande methode om dergelijke samenwerkingen te ontwerpen is uitgebreid met AI-agenten die zichzelf kunnen uitleggen. Verschillende van dergelijke ontwerpen zijn gepresenteerd (Hoofdstuk 7). Elk ontwerp varieerde in de mate van autonomie die de AI-agent had bij moreel gevoelige taken. Elk ontwerp gaf definieerde de rol van verschillende soorten uitleg die de AI-agent kan geven binnen dergelijke taken. Verschillende van deze voorgestelde ontwerpen zijn vervolgens geëvalueerd in het zorgdomein met eerstehulpverleners (Hoofdstuk 8). Uit de resultaten bleek dat deelnemers de uitleg waardeerden, maar ze deze ook als vervelend ervaarden wanneer ze tijdsdruk voelden. Bovendien voelden ze zich minder verantwoordelijk voor de AI-agent als deze autonomer werd. Dit verminderde hun motivatie om de geboden uitleg tot zich te nemen. Deze resultaten illustreert de complexiteit van het ontwerpen van een AI-agent en de uitleg die het geeft om de samenwerking met de mens te ondersteunen.

Bovenstaande bevindingen tonen dat uitleg van een AI-agent de potentie hebben om de samenwerking tussen mens en AI-agenten te verbeteren, aangezien dergelijke uitleg verschillende gunstige effecten kunnen hebben. Niet al deze effecten zijn echter positief. Verklaringen kunnen ook negatieve effecten hebben die nadelig zijn voor de samenwerking, wat grotendeels afhankelijk is van de context. Een overtuigender advies door een bijgaande uitleg zal nadelig zijn in een toepassing waar juist een kritische houding van de mens wenselijk is. Daarentegen is een dergelijke uitleg gunstig wanneer het ongerechtvaardigd wantrouwen mitigeert. De uitgevoerde studies toonden aan dat uitleg effecten met zich meebrengt die gewenst of ongewenst zijn naargelang de toepassing van de AI-agent. Deze en soortgelijke toekomstige studies die de verscheidenheid van effecten blootlegt die uitleg kan veroorzaken, zullen een solide basis vormen voor ontwerpkeuze hoe dergelijke uitleg ingezet moet worden. Dergelijke effecten en ontwerpkeuzes samengebracht kunnen worden in ontwerppatronen die de bevindingen en adviezen inzichtelijk maken gegeven een toepassing en samenwerkingsvorm tussen mens en AI-agent. Afgezien van deze inzichten, hebben we de waarde geïllustreerd om eigenschappen te formaliseren van actiegerichte uitleg, in plaats van uitleg met voornamelijk een epistemische waarde. Door een dergelijke formalisatie van eigenschappen en bijbehorende niveaus konden we een onderzoek agenda bieden voor dergelijke actiegericht uitleg. Zo kan doelgericht onderzoek verricht worden hoe uitleg menselijke autonomie in stand kan houden in interactie met AI-agenten die invloed uitoefent op diens leven. Ten slotte is de effectiviteit aangetoond van uitleg genererende methodes die onafhankelijk werken van de toegepaste machine-learning technieken die ten grondslag liggen van een AI-agent. Dit is met name het geval wanneer de uitleg niet elk detail van een AI-agent hoeft te onthullen. Bovendien zijn dergelijke methoden robuuster voor toekomstige ontwikkelingen in AI-onderzoek.

We sluiten af met een advies voor toekomstig onderzoek in XAI. Ten eerste is ons advies om meer aandacht te besteden aan de evaluatie van de effecten van uitleg en aan de theoretische onderbouwing hoe dergelijke uitleg kan worden gegenereerd. Er is behoefte aan meer rigoureuze evaluaties die gebaseerd zijn op realistische toepassingen van AI-agenten en expliciete theoretische modellen die zowel de positieve als negatieve effecten van uitleg beschrijven. Met meer aandacht voor de theoretische en wiskundige basis over hoe uitleg moeten worden gegenereerd, kunnen we bovendien toewerken naar methoden waarvoor we hard kunnen maken dat de gegenereerde uitleg voldoende accuraat is, ongeacht het intrinsiek complexe functioneren van een AI-agent. Met een verhoogde inspanning in beide, zouden we in staat moeten zijn om industrieën en overheden te voorzien van de kennis die nodig is om effectief en verantwoord uitleg toe te passen en correcte praktijken en regelgeving te formuleren. Ons tweede advies is om meer aandacht te besteden aan de rol en inbedding van uitleg in de mens-AI-samenwerking. Hierdoor worden de mogelijkheden van uitleg vergroot buiten alleen bij te dragen een betrouwbare AI-agenten. Zo kan uitleg langdurige samenwerkingen helpen ondersteunen, het bijdragen aan het ontdekken van nieuwe kennis, hulp bieden bij het opleiden van mensen in domein, en meer. Dit laatste advies komt neer op een grotere focus op mensgericht onderzoek over wat uitgelegd moet worden in plaats van welke uitleg gegeven kan worden.

CHAPTER 1

INTRODUCTION

We have started to create artificially intelligent (AI) agents with their own unique ways of reasoning and doing things. We adopt these AI agents in our lives, as they can effectively take over certain tasks and assist us in others. They can automate the living conditions in our homes, support the doctors that treat us, approve our loan requests, and determine our fit for our dream job, among many other tasks. We have come to notice the influence and impact AI agents have on the way we live our lives. Now, we must ensure that we are sufficiently involved and have a level of understanding of our AI agents appropriate to relinquish such influence to it.

As our AI agents become more independent, we risk them behaving in ways that violate our moral values and best intentions. For example, an AI agent played a significant role in the Dutch childcare benefits scandal, where Dutch tax authorities were unaware that the AI agent they used was wrongfully labelling people as fraudsters [1]. This example shows that AI agents can become harmful, such as a biased AI agent that discriminates against minorities or behaves strangely in a novel but critical situation. We want to be in control of our AI agents and hold humans responsible for their behaviour. The only way to achieve this control is to understand how, why, and when they behave in certain ways. This understanding is seen as the key to creating and using AI agents we can trust, rely upon, and collaborate with.

This need to understand has researchers invest in building AI agents capable of explaining themselves, a research field referred to as Explainable Artificial Intelligence (XAI). Ideally, we expect an AI agent to explain itself in the same way we expect a fellow human to explain their decision if that decision affects us in some way. Whether that decision is to adjust our home's thermostat, to determine our medical treatment or to approve or decline a new loan. In these cases, it would be ideal for the AI agent to explain its decision in a way that helps us accept, trust, and collaborate with it. In addition, an AI agent's explanations about itself can help us identify when an AI agent is harmful, how to remedy this harm and how to further improve the AI agent. Thus, an AI agent that can explain itself can offer numerous advantages.

There are two main challenges in achieving an AI agent as a competent explainer of itself: 1) identify what an AI agent should explain, and 2) enable the AI agent to generate such explanations. With only the former we might know what to explain for what purpose and what effects can be expected, but without an AI agent generating such explanations we cannot bring it into practice. With only the latter we might have AI agents capable of generating explanations, but without knowing what effects such explanations have and whether they fulfil a desirable purpose. Both challenges need to be addressed to ensure that Explainable AI agents can be responsibly applied in real-world applications. Hence, the main question XAI researchers face — and the topic of this thesis — is as follows:

What should an AI agent explain to humans,
and how can it generate such explanations?

By addressing this question, this thesis aims to contribute to the responsible application of AI agents by offering insights in what those agents should explain, what effects those explanations achieve, and how they can generate such explanations. With these insights we can design and develop Explainable AI agents with a purpose that match their intended real-world application.

1.1 The need for explainable artificial intelligence

Depending on whom you ask, XAI is either a new field or a resurgence of an older field. Historically, the combination of explanations and AI is not novel. Decades ago, when AI research focused on what is now sometimes referred to as “good old-fashioned AI” (GOFAI), explanations were used to communicate knowledge within expert systems to the humans collaborating with them [2, 3, 4]. These explanations had an educational role, with only a few exceptions aiming to explain how expert systems reasoned. This approach was mostly due to expert systems reasoning about elicited human knowledge in a way that is already familiar to us, meaning no explanations were needed to explain this reasoning [5]. In contrast, current AI agents are different from AI agents of previous generations. The focus now lies on learning from data as opposed to reasoning about knowledge [6]. The role of explanations has shifted from educational to translational, making sense of the profoundly alien way a present-day AI makes decisions [7, 8].

Whether XAI is a novel field or simply a resurgence can be debated, but either way the need for XAI is more apparent than ever. The most effective AI of today is not engineered but taught. Their intelligence emerges from the use of iterative algorithms applied to massive amounts of data using unprecedented computational power [9]. This emergence of intelligence based on a complex learning process results in AI agents that are difficult to understand [10]. We have crossed a threshold, now being able to build highly effective AI agents that can surpass humans on a variety of tasks. Such AI agents tend to consist of billions of learned parameters. No human can comprehend how those parameters interact and how this interaction results in their impressive performance. XAI aims to offer much-needed understanding of how to use AI without worry of its consequences. XAI seeks how to facilitate the understanding that allows us to determine whether and when an AI can be trusted and relied upon [7].

European political bodies propose to regulate AI agents and to set their ability to explain itself as a requirement. For example, consider the right to explanation defined in the European Union's GDPR from 2016 [11]. This loosely defined "right" has since been further substantiated in a proposal for a legal framework to regulate AI agents [12]. This framework dictates that the high-risk AI agents which affect our lives "shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system's output and use it appropriately" (Article 13, 1st paragraph). This regulation underlines the need for humans to understand an AI agent to improve their collaboration with it.

Such regulations sadly originate from cases where AI agents caused harm. In recent years, the lack of understanding in our AI agents has caused negative societal impact in some high-profile cases. For example, consider the AI agent showing racist behaviour that was used to predict recidivism risk in the U.S. justice system [13], the AI agent from Google that tagged people on photos as gorillas based on their ethnicity [14], or the AI agent at Amazon who filtered out applicants for certain vacancies based on gender [15]. In the Netherlands, we face the consequences of the Dutch childcare benefits scandal, which involved the use of an AI agent that discriminated against parents based on their dual citizenship [1].

The future of AI can be brighter with the advance of the field of XAI. The agreed upon hypothesis in the XAI community is that with an AI agent capable of explaining its decisions, a harmful AI agent can be more easily identified, prevented, or otherwise mitigated [7]. The research community is working hard on solutions, while industry is quickly adopting them [16]. Aside from the verification and corrections that XAI enables for an AI agent, researchers in this field agree that explanations improve the acceptance and effective use of them as well [17]. This improved acceptance is particularly relevant for domains that have been historically hesitant to accept the use of an AI agent [18]. These domains include both healthcare and the military, where decisions can have major negative effects and where our society tends to be conservative in introducing AI agents. The potential for XAI research and its results to improve the acceptance and use of AI agents thus further substantiates the economic and societal benefits of the research.

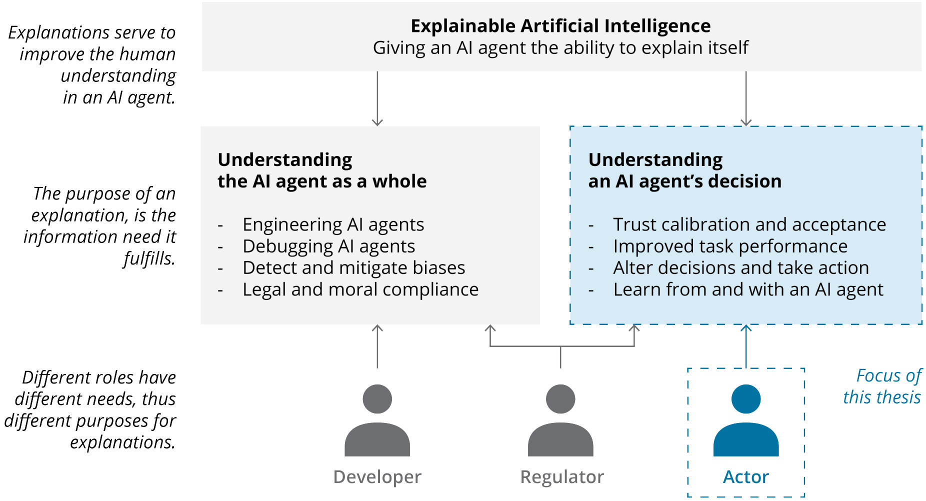

Explanations serve to improve the human understanding in an AI agent.
The purpose of an explanation, is the information need it fulfills.
Different roles have different needs, thus different purposes for explanations.
Explainable Artificial Intelligence
Giving an AI agent the ability to explain itself
Understanding the AI agent as a whole
- Engineering AI agents
- Debugging AI agents
- Detect and mitigate biases
- Legal and moral compliance
Understanding an AI agent's decision
- Trust calibration and acceptance
- Improved task performance
- Alter decisions and take action
- Learn from and with an AI agent
Developer
Regulator
Actor
Focus of this thesis

Figure 1.1: An overview of the main purposes of explanations addressed by the Explainable AI community. The dashed and highlighted components receive the focus in this thesis.

In sum, there is a societal and economical need for AI agents to explain their decision making, which the XAI research field tries to fulfil. Before we detail our research's efforts to advance the state of the art in XAI in the following chapters, we first elaborate on the current state of XAI.

1.1.1 The many purposes of explanations

All explanations aim to improve the collaboration between human and AI agents on a joint task, however different explanations can have different purposes within that collaboration [19]. In Figure 1.1 we summarise the most noted purposes of an AI agent's explanation in the literature.

We distinguish between three human roles relative to the AI agent, each imposing their own primary purposes: 1) the developer, 2) the regulator and 3) the actor. These roles derive from those identified by Arrieta et al. [20], Ribera et al. [21] and Greeff et al. [22]. The developer helps create, deploy, and update an AI agent. The regulator decides when and how an AI agent is used and whether its application remains within the appropriate legal and moral frameworks. Finally, the actor is the human directly dealing or collaborating with the AI agent. The actor can have full autonomy, where the AI offers only advice or suggestions (e.g., on a medical diagnosis); limited autonomy, where the AI agent dictates a decision (e.g., it declines your loan application); or anything in between (e.g., some decisions left to the human others to the AI agent). Relevant actors include clients, patients, operators, or consultants. Notably, these roles are not fixed. For instance, an actor whose loan application is declined can also serve as a regulator if a decision feels unjust to that actor. Similarly, a regulator might assume the role of developer if it can provide direct feedback to the AI agent it will use to improve itself.

Aside from distinct roles, we differentiate between two categories of understanding: 1) an objectual or global understanding which is the understanding how an AI agent functions as a whole [23] or 2) a post-hoc understanding which is the understanding of how an AI agent came to a decision [24]. These two categories are not mutually exclusive. Many subsequent post-hoc explanations might induce an objectual understanding of an AI agent (e.g., an understanding how decision trees function). Similarly, a sufficiently detailed objectual understanding can help indicate why a single decision was made (e.g., which decision rules played a role in a prediction). However, the category of understanding between the two differ, as do the purposes for such understanding.

We use these three roles and two categories to provide an overview of the most common purposes for explanations, illustrated in Figure 1.1. The developer is mostly interested in an objectual understanding. Objectual explanations can support the debugging of an AI agent [25, 26], detecting whether an AI agent is biased [16, 27], and how compliant an AI agent is to legislation [28]. By contrast, the actor is more interested in a post-hoc understanding of a decision, wanting to know whether a decision can be trusted and accepted [7] or altered [29], or how the AI agent improves joint task performance [30] and learn co-actively to perform such a task [31]. The regulator has the unique position of needing both a limited objectual and post-hoc understanding. As the regulator wants to both verify and evaluate an AI agent as a whole, but also to understand specific decisions as necessary [20, 22].

The above-mentioned purposes have so far received the most attention in XAI. For a more in-depth overview of these and other purposes, we refer to Lipton [32], Doshi-Velez and Kim [33], Samek et al. [34], Abdul et. al [35], Gilpin et al. [36] and Herman et. al [37]. Within this thesis, we focus on the actor role and their need for a post-hoc understanding of singular decisions. These explanations are deemed most useful for humans who adopt AI agents in their everyday lives [18].

1.1.2 An explanation’s class, form, modality, and method

In the past years, hundreds of methods to generate explanations have been developed. This need is reflected by the number of articles reviewing such methods. For instance, take the reviews from Adadi et. al [38], Guidotti et. al [39], Arrieta et. al [20] and Linardatos et. al [40], these reviews combined report on 303 unique methods. The extensive review from Nauta et. al [41] report that between 2016 and 2020 a total of 361 novel methods have been published, with more than half that number (167) published in 2020, showing an exponential increase in methods. Each review tends to adopt a different taxonomy to categorize XAI research, mostly due to lack of consensus in the field's terminology [32]. Within this thesis, we do not attempt to provide a new or even complete taxonomy, though for reference we use the terms "explanation class", "explanation form", "explanation modality" and "generative method".

An explanation class is the explanation's conveyed information and, thus, relates to the question it answers or human information need it fulfils. Reviews such those from Adadi et al. [38], Guidotti et al. [39] and Arrieta et al. [20] provide comprehensive overviews of various explanation classes. For this thesis, the important classes are feature attributions [17], contrastive explanations [42], confidence explanations [43], and the novel class of actionable explanations. Feature attributions and contrastive explanations comprise the state of the art of explanations, actionable explanations are increasingly cited as necessary but research on them is limited, and confidence explanations are proposed in this thesis as a novel class. A feature attribution depicts which features were the main causes of an AI agent's decision. Contrastive explanations regard why one decision was made instead of another. An actionable explanation explains what changes are required for the AI agent to decide differently and how to implement such changes. Confidence explanations aim to compute and explain how likely a decision is to prove correct.

Common explanation forms include feature-based [44], example-based [45], and rule-based [46]. These forms, respectively, provide feature values and distributions, specific behavioural examples, and decision rules for describing an AI agent's decision making and reasoning. Although other forms exist, in this thesis we limit ourselves to feature-, example-, and rule-based forms. These forms have proven themselves as comprehensible by humans and can be generated through a variety of approaches.

The modality of an explanation captures how the information is communicated. Common explanation modalities are a visual but static communication of the explanation [47], communication through interactive interface components and dashboards [48], or through text and dialogues [49]. Visualizations include bar graphs, tornado diagrams, and feature highlights [35]. Such static visualizations contrast with the more interactive interfaces and dashboards that allow for the creation of one's own visualization or to support a progressive discovery of knowledge. Recently, textual and dialogue-based explanations are being explored as a more natural modality for communicating explanations [50]. Examples include the automatic generation of explanatory texts, the development of chat bots and interactions through avatars with dialogue capabilities.

Explanation generating methods, or simply methods, are the algorithms and techniques used to obtain the necessary information from an AI agent. We differentiate between the use of surrogate models or intrinsically interpretable models [51]. A surrogate model approximates the AI agent's decision making and supports the extraction of the necessary information to generate an explanation (i.e., a form of reverse engineering and model inference) [17]. These methods are often used when the AI agent is deemed too complex to obtain the required information directly (e.g., with deep neural networks) or when the method is intended to apply to any AI agent (e.g., by only assuming input-output access). The use of surrogate models is opposed to the use of intrinsically which aim to result in an AI agent that offers direct access to the required information. For instance, the CART algorithm allows the creation of a decision tree based on data that an AI agent can use to base its decision on [52]. Decision trees tend to be more interpretable whose content can be easily accessed to base explanations on [53].

Typically, a trade-off between an AI agent's intrinsic interpretability and its performance is assumed [7]. For instance, a deep neural network is viewed as achieving high performance at the cost of interpretability, whereas a decision tree is seen as limited in performance but more interpretable. However, such statements do not necessarily hold in practice [58], nor can that distinction always be made. For instance, a small neural network might be more interpretable than a large decision tree. Similarly, given the impact someone's background has on the explainer's decision, a regression model might be more interpretable according to a social scientist due to their familiarity with applied statistics than is a support vector machine that is much more mathematical in nature. Neither model might be interpretable for a layperson. The background and expertise of a human has a significant impact on the design of an explanation. The degree to which an AI agent is interpretable dictates how easy it is to access the necessary information to generate a required explanation given a human's background and information need. As long as we do not know what explanations are required and for whom, a debate on the trade-off between performance and interpretability is meaningless. Without knowing what needs to be explained, it cannot be determined what approach is required to generate the explanations – i.e., the use of intrinsic interpretable or surrogate models.

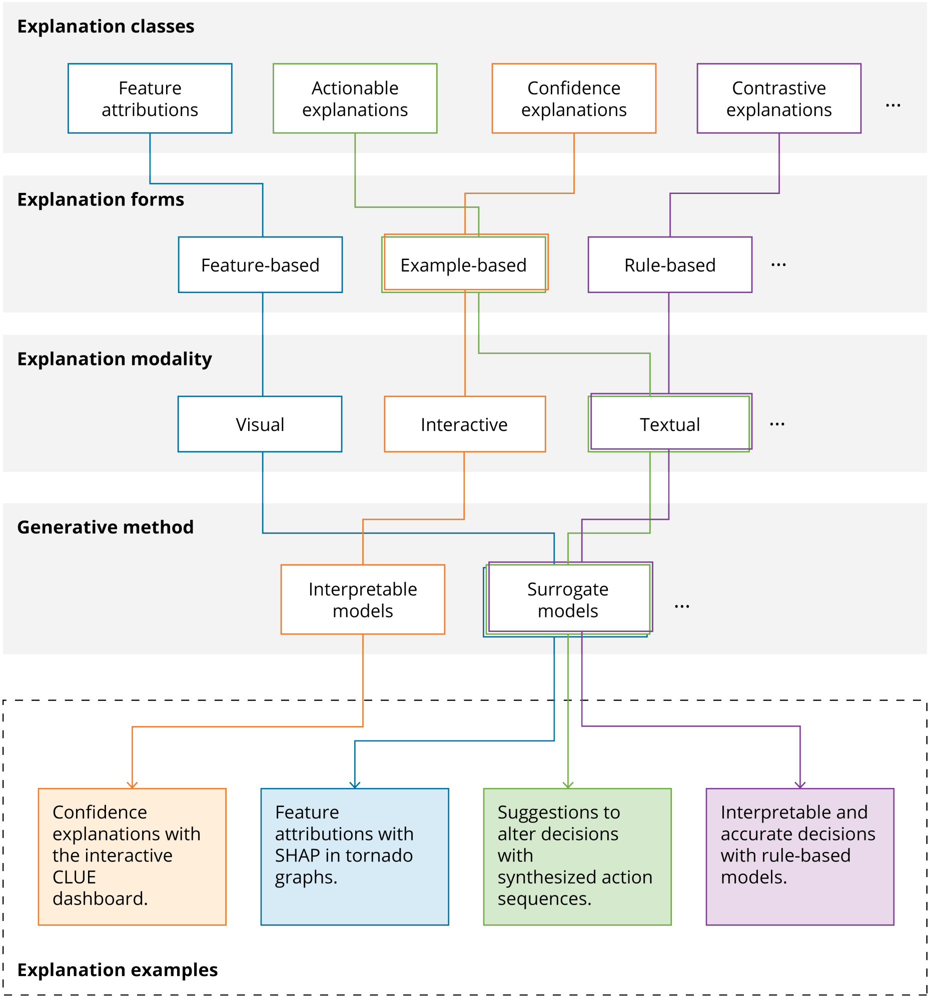


Figure 1.2: An overview of several explanation classes, forms, modalities, and methods to generate them. Combined they form an implementation of an explanation. Depicted here are four examples; CLUE [54], SHAP [55], Synthesized Action Sequences [56], and Falling Rule List models [57].

Only the combination of an explanation class, form, and modality together with a method to generate it results in an actual explanation. A method only extracts the information for a specific class of explanations, the form dictates the shape of that information, with the modality dictating how it is communicated. See Figure 1.2 for an illustration of four examples of such combinations: the dashboard named CLUE for interactive explanations of an AI agent's confidence [54], the feature-attribution method called SHAP visualized in tornado graphs [55], action suggestions to alter an AI agent's decision through synthesized action sequences [56] and explanation of why a decision was made through rule-based models [57].

1.1.3 The fragmented field of XAI

With the rapid technological progress in AI and the associated and impressive results, the field of XAI is equally governed by a technology-centred perspective [33]. Most of the publications within XAI propose novel methods of generating explanations [38, 39]. These methods tend to lack evaluation aside from an estimate of their accuracy and computational efficiency. Nauta et. al [41] report that between 2016 and 2020 a total of 361 papers proposed novel methods whose evaluations were limited to such benchmarks. Only 49 publications in that same period reported on a more involved evaluation of the generated explanation. Doshi-Velez and Kim [33] further state that the field of XAI follows the notion of "you'll know it when you see it" (p1). Authors of novel methods seem to rely on their own intuition to determine if the explanation their method generates is beneficial and valuable [59].

Several scholars criticize this technology-centred perspective towards XAI research within its community and advocate for a more human-centred perspective. Some argue that XAI should focus more on rigorous evaluations to assess the effects of explanations [60]. Doshi-Velez and Kim [33] define three distinct levels of evaluations: 1) application-grounded evaluations, 2) human-grounded evaluations, and 3) functionally-grounded evaluations.

Application-grounded evaluations involve an explanation's evaluation in a realistic context, uses an actual AI agent, and a representative population sample that fits the context. These evaluations provide detailed results on the effects of an explanation in a particular application. Human-grounded evaluations are suited to test the general notions of an explanations. These tend to use a simplified AI agent, an approximation of a relevant context, laypeople as participants or a combination thereof. The functionally-grounded evaluations make use of metrics serving as proxies of an explanation's quality that can be applied to the explanations generated by some methods. These do not include human participants and function to evaluate a method's adherence to an explanation class that was either evaluated previously or cannot be evaluated due to practical or ethical concerns. Note that the first two levels, the application- and human-grounded evaluations, can be done quantitatively to statistically show an explanation's effect or qualitatively to explore the potential effects of an explanation. Within this thesis we make use of both quantitative and qualitative human-grounded evaluations and functional evaluations of developed methods. No application-grounded evaluations are reported in this thesis.

The criticism towards the technology-centred perspective is that it invites solely a focus on functionally-grounded evaluations with little or no reference to the human or application-grounded applications [33]. Going a step further, others opposing the technology-centred perspective argue that such a perspective will result in the irresponsible use of methods in real-world applications [58, 23]. The common argument here is that, in practice, explanations require human-grounded evaluations to give direction what explanations might be beneficial for some purpose, followed by detailed application-grounded evaluations to further substantiate those findings. Only then can the explanation be responsibly applied within an application.

Aside from rigorous evaluations, one can also rely on a sound theoretical foundation based on human psychology to determine what explanations should be used for what purpose and in what kind of application [61, 59]. Though ideally, one uses such a foundation to design an explanation, then evaluates the assumptions and hypotheses made through human-grounded evaluations, followed by application-grounded evaluations to validate the findings.

Finally, there are those who oppose the use of a self-explaining AI agent entirely [62, 63, 64]. These criticisms do not necessarily discredit the value of explanations, instead they argue that more is needed to achieve trustworthy and responsible AI agents. They argue that an AI agent can be developed to omit the need for it to explain itself, for instance by having accredited development parties, providing clear documentation, educating those interacting with the AI agent, and appropriate design processes to support that interaction. This stance omits the added value explanations offer to the collaboration between humans and AI agents. This thesis agrees that to achieve trustworthy AI more is needed than explanations, though explanations form a part of the solution. Even when the AI agent went through rigorous testing, has enough documentation and the humans collaborating with it are properly educated, explanations still have the potential to improve and enrich this collaboration.

Another trend, associated to an overall technological-centred perspective, is that many developed methods with a lacking rigorous evaluation, are often made open-source. The industry is quick to adopt these open-source methods to accommodate the (expected) legal, ethical, and economical requirements for AI agents [65]. We risk disastrous effects to our society with this trend to adopt methods and a lack of evaluations of the effects the explanations from these methods bring about [58]. We run the risk that our society presumes an under-

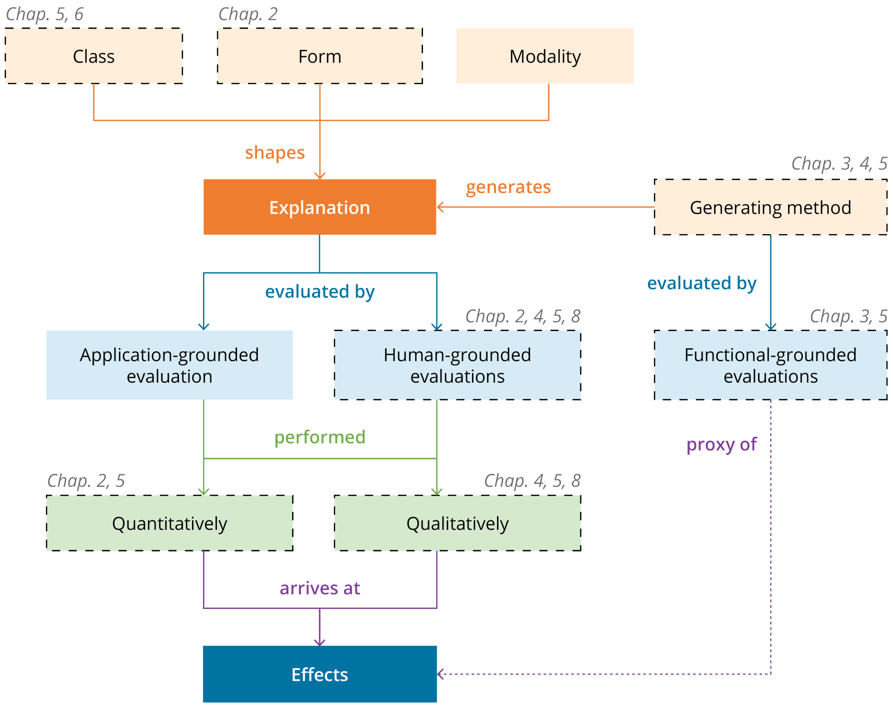


Figure 1.3: An overview of the aspects an explanation consists of (class, form, modality, and method) and the ways of evaluating such explanations. The encircled components receive the focus in this thesis.

standing of each AI agent used, while the explanations may in fact have only a small positive or even negative, effect on our actual collaboration with them [66, 67].

At the same time, the technology-centred community seems to refer increasingly to the human-centred perspective to XAI (e.g., see the ever-increasing number of references to position papers from Miller et al. [59, 68]). Such papers receive hundreds of references each year, and most authors who refer to them introduce novel explanation methods. The question remains, however, to what extent these methods are generalizable, and how well their effects are known through rigorous evaluations before being put to practice.

Figure 1.3 illustrates the focus of this thesis given the types of evaluations and the explanation's class, form, modality, and generative method.

1.2 Research approach and aim

In this thesis, we adopt both the human- and technology-centred perspective to XAI research. We limit ourselves to four purposes of explanations identified by reviewing the human-centred research on AI and XAI. The first is the explanation purpose of offering laypeople an understanding in why the AI agent makes a certain decision. The second is calibrating the human's trust in an AI agent's decision. The third is the human's ability to contest those decisions effectively. The fourth purpose we address is for explanations to improve the collaboration between human and AI agent, particularly in high-risk and morally sensitive tasks. We identified these purposes by reviewing the human-centred literature on AI and XAI. Furthermore, we evaluate our designed explanations using methods inspired or taken from that same literature. Finally, if there were no methods to generate our designed explanations, we developed our own in most cases based on approaches taken from the technology-centred literature on XAI.

We argue that only with a combined human- and technology-centred perspective can we arrive at explanations that can be responsibly and feasibly applied in real-world applications. In this thesis in particular, the societal need for XAI made us focus on applications where the decisions of AI agents directly impact the lives of the involved actors (e.g., the expert working with the AI agent or the human eventually subjected to the decisions made). In summary, our research aim is as follows:

> To design and develop explanations that support a responsible
> and effective collaboration between humans and AI agents.

1.3 Research questions, hypotheses, and outline

This thesis consists of three parts, each addressing an aspect of the above described research aim. The first part addresses the contrastive explanation class that aims to convey why the AI agent made a certain decision opposed to another. This part addresses what form this class should take and how such explanations can be generated. The second part introduces two novel explanation classes and defines their properties: confidence explanations and actionable explanations. In the case of confidence explanations, we propose a method to generate them. For the case of actionable explanations, we provide a research agenda to address such explanations further based on current research gaps.

The third part addresses the effects of various explanation classes, forms, and modalities on the collaboration between human and AI agents. A collaboration in high-risk and morally sensitive tasks. With each consecutive part, we address our research aim towards the entire human–AI collaboration. Below, we discuss each part in more detail, followed by the thesis outline.

1.3.1 Part I: Explanations for understanding

Part I addresses how an AI agent should explain the reasons it made one decision instead of another, namely the contrastive explanation.

The oft-referenced purpose of an AI explanation is to induce human understanding in the AI agent's decision-making process [17]. Contrastive explanations aim to induce such an understanding, as they address most "Why?" questions humans have [68]. A sub-class of these explanations are the counterfactual explanations which convey the minimal changes for an AI agent to behave differently [29]. Contrastive explanations induce an exact understanding of an AI agent's decision boundaries [42]. On the other hand, the similar explanations of feature attributions induce only an understanding of which features are central to decisions. A feature attribution provides an explanation as, "feature x is most important for this decision", whereas a (rule-based) contrastive explanation provides an explanation as, "because feature x is above threshold t, this decision was made instead of another".

It could be beneficial if any AI could provide such a contrastive explanation. Their contrastive nature, comparing the reasons for making one decision to the reasons for making another, naturally limits the explained information to the key differences of interest to a human [68]. However, what form the information in a contrastive explanation should take is not yet understood, nor how that information can be extracted from any given AI agent. In this part, we explore whether an example- or rule-based form is the most effective for a contrastive explanation class. In addition, we explore how such an explanation can be generated for two AI technologies; classification models used in decision-support tools and reinforcement learning agents used in autonomous systems. This approach allows us to assess the effectiveness and feasibility of generating contrastive explanations for various kinds of AI agents.

The associated research questions in Part I are:

RQ 1: How should an AI agent explain why it made one decision instead of another?

RQ 1.1: What effect do example- and rule-based contrastive explanations have on the understanding of an AI agent's decision making?
- *Chapter 2*

RQ 1.2: Which explanation generating method allows for the generation of rule-based contrastive explanations for classification models used in decision support tools?
- *Chapter 3*

RQ 1.3: Which explanation generating method allows for the generation of example-based contrastive explanations for reinforcement learning agents used in autonomous systems?
- *Chapter 4*

1.3.2 Part II: Explanations to act upon

Part II addresses how explanations from an AI agent help humans determine when and how to act on the agent's decision.

A post-hoc understanding on how an AI agent decides is viewed by XAI researchers as a step towards knowing how to collaborate with an AI agent, as it is expected that such understanding leads to better calibrated trust, which in turn should improve joint task performance [61]. Such an understanding allows humans to infer how they can contest and alter an AI agent's decision, particularly when they are displeased with the decision [29]. Instead of relying solely on explanations to induce a post-hoc understanding in the hope of calibrating trust and support contestability, we argue that explanations can do more. In fact, humans adjust their explanations for their intended purposes [69], so why not design an AI agent's explanations for such purposes directly? An AI agent can achieve more with its explanation than simply induce an understanding of its function [23].

In this second part, we focus on two explanation classes. First, we introduce the novel class of confidence explanations. Second, we formally define the up to now ambiguous class of actionable explanations. Confidence explanations aim to calibrate one's trust in and reliance on an AI agent's decision. We introduce this class, define its required properties, validate them, and propose a concrete methodology both to compute confidence and to explain this computation. An actionable explanation aims not only to induce post-hoc understanding but also to support humans in inferring how an AI agent's decision can be altered. We review the literature on this explanation class and formally define the properties required to make an explanation actionable. In addition, we propose a research agenda based on the identified research gaps according to a literature review. With both explanation classes we aim to support humans in deciding whether and how to act on an AI agent's decision, informing the human on its trustworthiness and reliability and how to contest and alter the decision effectively.

The following research questions are addressed in Part II:

RQ 2: Which classes of explanations enable humans to decide whether and how to act on an AI agent's decision?

RQ 2.1: How should an AI agent compute and explain its confidence such that humans can decide when to trust and rely upon the agent's decision?
- *Chapter 5*

RQ 2.2: What explanation properties support humans in effectively altering an AI agent's decision to make it more favourable?
- *Chapter 6*

1.3.3 Part III: Explanations in human–AI collaboration

In Part III we propose a design methodology how explanations can be responsibly integrated into human–AI collaborations and evaluate several collaboration designs.

Johnson and Alonso [70] said “No AI agent is an island”, meaning that for any AI agent to be successful, it needs to collaborate with humans. Currently, even the literature from the human-centred perspective within XAI does little to address the use of explanations to explicitly support human–AI collaboration [71]. The first two parts of this thesis have a similarly narrow focus. Hence, in this third part, we abstract from specific explanation classes and address the role of explanations in the context of human–AI collaboration. Specifically, we look towards joint tasks with a morally sensitive element, which make for high-risk applications of AI agents.

We first propose a design method that allows for the embedding of explanations in human–AI collaboration in morally sensitive tasks and their moral context. This method is demonstrated through several common collaboration designs. We extend each design with the idea of explaining moral context to support making joint decisions. Through a qualitative experiment involving medical domain experts, we then evaluate several collaboration designs and explore the effects of various explanation classes in each.

In this Part III, we aim to answer the following research questions:

RQ 3: What is the role of explanations when human and AI agents collaborate on morally sensitive tasks?

RQ 3.1: What is a suitable design method for human–AI collaborations that responsibly incorporates explanations of the moral context?
- *Chapter 7*

RQ 3.2: What are the effects and functions of explanations according to domain experts collaborating with an AI agent on a morally sensitive task?
- *Chapter 8*

1.3.4 Outline

Part I addresses how an AI agent can explain why it made a decision. Part II addresses which explanations aid the human in determining how to act upon an AI agent's decision. Part III addresses the design, function, and effects of explanations in a human–AI collaboration on morally sensitive tasks. Finally, Part IV discusses the findings, the limitations, and the societal impact of this research. It also provides concluding remarks and future work and advice on the direction of the field of XAI.

This outline is illustrated at the chapter level in Figure 1.4. This figure highlights for each chapter the class, form, modality, and method worked on if applicable. It also addresses the kind of evaluation performed: human- or functionally-grounded; qualitatively or quantitatively. No application-grounded evaluations were performed.

Each chapter consists of a published journal or conference paper with minimal changes. To prevent overlap with this introductory chapter, these changes constitute the adaptation of the abstract and the shortening of chapter introductions. Furthermore, minor changes were implemented in most chapters to ensure consistent use of terms and writing style, as introduced in this chapter. These changes are outlined on the first page of each chapter.

1.4 Defining the Scope

The scope of this research is determined by how we interpret the key concepts, made choices and assumptions:

- We define an "explanation" in general as a communication act from one agent to another containing one or more statements that clarify an event, context, or process. Within the context of this research, we define an "explanation" as the clarification an AI agent provides a human about its internal process and related aspects such as its observations or expected consequences of a decision.

- We do not address the modality of explanations, only the classes, forms, and methods.

- We do not account for explanations given by the human that the AI agent should be able to interpret and learn from. In addition, our explanations serve to disclose how the AI agent functions, performs, or behaves. We do not explicitly aim to explain how to conduct the task nor to teach about the domain or situational context.

- When we refer to "human", we typically refer to humans with the role of an actor, as previously explained. This human might be a domain expert or a layperson, depending on the use case discussed; however, they are never experts in AI agents. We do not address explanations that are beneficial only to humans with the developer or regulator roles. Certain results might apply to such roles as well, although this supposition is never validated in this thesis and is left for future work.

- Throughout, it is assumed that the human involved is motivated to collaborate with the AI agent and does so knowingly. An exception to this is Chapter 6 where the human is subjected to an AI agent's decision and not necessarily aware of this nor an expert in the domain. Furthermore, we do not consider more specific characteristics such as (digital) literacy and the requirements such characteristics set upon explanations.

- When we refer to an “AI agent”, we typically refer to a software system capable of making decisions based on its perceptions of a situation. A decision is assumed to be implemented either by the AI agent itself or by the human to which some decision was offered as advice.

- The explanations in this thesis are all post-hoc, meaning that they aim to explain a single decision. This is opposed to objectual explanations that aim to explain the entire AI agent and its origin.

- There are multiple ways to implement an AI agent. We attempt to abstract from any specific implementation and aim for explanations that can be generated in a model-agnostic way. Model-agnostic implies that only input-output access to an AI agent is assumed, making no further assumptions on how the AI agent functions internally. However, we assume in general that an AI agent is build using one or more machine learning models and associated components.

Figure 1.4 (next page): The outline of this thesis. For each chapter, a brief description is given; the explanation’s class, form, modality, and method where applicable; and the type of evaluation performed. The relevant aspect or evaluation that receives the focus in that chapter is underlined.

Summary
The executive summary of this thesis

Introduction
An introduction to the field of XAI, this thesis' research questions, its aim and scope.

Part I - Explanations for understanding

Chapter 2
Rule- or example-based?

Class:	Contrastive explanations
Form:	Rule-based, example-based
Modality:	Textual, visual
Method:	N/A
Evaluation:	Quantitative human grounded

Chapter 3
How to generate contrastive explanations for decision support systems?

Class:	Contrastive explanations
Form:	Rule-based
Modality:	Textual
Method:	Surrogate model
Evaluation:	Functionally grounded

Chapter 4
How to generate contrastive explanations for autonomous systems?

Class:	Contrastive explanations
Form:	Example-based
Modality:	Textual
Method:	Instrinsically interpretable
Evaluation:	Qualitative human grounded

Part II - Explanations to act upon

Chapter 5
A new class: confidence explanations.

Class:	Confidence explanations
Form:	Example-based
Modality:	Textual
Method:	Surrogate model
Evaluation:	Quantitative and qualitative human grounded; functional grounded.

Chapter 6
Defining a class: Actionable explanations.

Class:	Actionable explanations
Form:	Example-based
Modality:	N/A
Method:	N/A
Evaluation:	N/A

Part III - Explanations in human-AI collaboration

Chapter 7
A method to design human-AI collaborations with explanations.

Class:	Contrastive and confidence explanations; feature attributions
Form:	Rule-, example and feature-based
Modality:	N/A
Method:	N/A
Evaluation:	N/A

Chapter 8
What is the role of explanations in a collaboration?

Class:	Contrastive and confidence explanations; feature attributions
Form:	Rule-, example and feature-based
Modality:	Textual, Visual
Method:	N/A
Evaluation:	Qualitatively human grounded

Part IV - Conclusions and discussion

Chapter 9
Discussion of the findings, limitations and the societal implications of the research.

Chapter 10
Conclusions and directions for future work.

PART I

EXPLANATIONS FOR UNDERSTANDING

CHAPTER 2

CONTRASTIVE EXPLANATIONS WITH RULES AND EXAMPLES

This chapter is adapted from; **van der Waa, J.**, Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291:103404*. The adaptations include an altered abstract, shortened introduction, and an adjusted lay-out, including the formatting of figures and tables.

Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, Mark Neerincx

In this chapter, we evaluate two explanation forms. Often, explanation generating methods are based on face-value notions of what constitutes an effective explanation form instead of being based on findings from human evaluations. Two examples of such explanations are contrastive rule- and example-based explanations. We evaluate the effects of these two forms of contrastive explanations on AI understanding, persuasive power, and task performance in the context of decision support for diabetes self-management. Furthermore, we provide three sets of recommendations based on our experience designing this evaluation to help improve future evaluations. These address the need for a motivated theoretical framework of expected effects between constructs, the selection and use of the task and context, and the selection of appropriate measurements for key constructs. Our results show that rule-based explanations slightly improve the user's understanding of the AI agent's decision making, whereas both rule- and example-based explanations persuade humans to follow the advice even when incorrect. This applies especially for the example-based explanations. Neither explanation improved task performance compared to no explanation. These findings may be explained by both explanation forms providing superficial details only without further interaction. They might not convey an underlying rationale or causality from which the participants could construct a generalized mental model on how the AI agent made decisions.

2.1 Introduction

Humans expect others to comprehensibly explain decisions that have an impact on them [72]. The same holds for humans interacting with decision support systems (DSS). To help them understand and trust a system's reasoning, such systems need to explain their advice to humans [72, 68]. Currently, several approaches are proposed in the field of Explainable AI (XAI) that allow DSS to generate explanations [39]. Aside from the numerous computational evaluations of implemented methods, literature reviews show that there is an overall lack of high quality human evaluations that add a human-centered perspective to the field of XAI [33, 73]. As explanations fulfil a human need, an evaluation is needed whether these explanations fulfil that need. This can provide valuable insights into an explanation's requirements and effects. In addition, evaluations can be used to benchmark XAI methods to measure the research field's progress.

The contribution of this chapter is twofold. First, we propose a set of *recommendations* on designing human evaluations in the field of XAI. Second, we performed an extensive evaluation on the effects of *rule-based* and *example-based* contrastive explanations. The recommendations regard 1) how to construct a theory of the effects that explanations are expected to have, 2) how to select a use case and participants to evaluate that theory, and 3) which types of measurements to use for the theorized effects. These recommendations are intended as a reference for XAI researchers unfamiliar to human evaluations. They are based on our experience designing a human evaluation and retread knowledge that is more common in fields such as cognitive psychology and Human-Computer Interaction.

The presented evaluation focuses on two forms of contrastive explanations and their evaluation. Contrastive explanations in the context of a DSS are those that answer questions as 'Why this advice instead of that advice?' [42]. These explanations help humans to understand and pinpoint information that caused the system to give one advice over the other. In two separate experiments, we evaluated two contrastive explanation forms. An explanation form defines the way information is structured and is often defined by the algorithmic approach to generate explanations. Note that this is different from explanation form, which defines how it is presented (e.g. textually or visually). The two evaluated forms were *rule-based* and *example-based* explanations with no explanation as a control. These two explanation forms are often referred to as means to convey a system's internal workings to a human. However, these statements are not yet formalized into a theory nor compared in detail. Hence, our second contribution is the evaluation of the effects that *rule-based* and *example-based* explanations have on *system understanding* (Experiment I), *persuasive power* and *task performance* (Experiment II). We define system understanding as the human's ability to know how the system behaves in a novel situation and why. The persuasive power of an explanation is defined as its capacity to convince the human to follow the given advice independent of whether it is correct or not. Task performance is defined as the decision accuracy of the combination of the system, explanation and human. Together, these concepts relate to the broader concept of trust, an important topic in XAI research. System understanding is believed to help humans achieve an appropriate level of trust in a DSS, and both system understanding and appropriate trust are assumed to improve task performance [61]. Explanations might also persuade the human to various extents, resulting in either appropriate, over- or under-trust, which could affect task performance [74]. Instead of measuring trust directly, we opted for measuring the intermediate variables of understanding and persuasion to better understand how these affect the task.

The way of structuring explanatory information differs between the two explanation forms examined in this study. *Rule-based* explanations are "if ... then ..." statements, whereas *example-based* explanations provide historical situations similar to the current situation. In our experiments, both explanation forms were *contrastive*, comparing a given advice to an alternative advice that was not given. The *rule-based contrastive explanations* explicitly conveyed the DSS's decision boundary between the given advice and the alternative advice. Whilst the *example-based contrastive explanations* provided two examples, one on either side of this decision boundary, both as similar as possible to the current situation. The first example illustrated a situation where the given advice proved to be correct, and the second example showed a different situation where an alternative advice was correct.

Rule-based explanations explicitly state the DSS's decision boundary between the given and the contrasting advice. Given this fact, we hypothesized that these explanations improve a participant's *understanding* of system behaviour causing an improved *task performance* compared to *example-based* explanations. Specifically, we expected participants to be able to identify the most important feature used by the DSS in a given situation, replicate this feature's relevant decision thresholds and use this knowledge to predict the DSS's behaviour in novel situations. When the human is confronted with how correct its decisions were, this knowledge would result in a better estimate when a DSS's advice is correct or not. However, *rule-based* explanations are very factual and provide little information to convince the participant of the correctness of a given advice. As such, we expected *rule-based* explanations to have little *persuasive power*. For the *example-based* explanations we hypothesized opposite effects. As examples of correct past behaviour would incite confidence in a given advice, we hypothesized them to hold more *persuasive power*. However, the amount of *understanding* a participant would gain would be limited, as it would rely on participants inferring the separating decision boundary between the examples rather than having it presented to them. Whether persuasive power is desirable in an explanation depends on the use case as well as the performance of the DSS. A low performance DSS combined with a highly persuasive explanation for example, would likely result in a low task performance.

The use case of the human evaluation was based on a diabetes mellitus type 1 (DMT1) self-management context, where patients are assisted by a personalized DSS to decide on the correct dosage of insulin. Insulin is a hormone that DMT1 patients have to administer to prevent the negative effects of the disturbed blood glucose regulation associated with this condition. The dose is highly personal and context dependent, and an incorrect dose can cause the patient short- or long-term harm. The purpose of the DSS's advice is to minimize these adverse effects. This use case was selected for two reasons. Firstly, AI is increasingly more often used in DMT1 self-management [75, 76, 77]. Therefore, the results are relevant for research on DSS aided DMT1 self-management. Secondly, this use case was both understandable and motivating for healthy participants without any experience with DMT1. Because DMT1 patients would have potentially confounding experience with insulin administration or certain biases, we recruited healthy participants that imagined themselves in the situation of a DMT1 patient. Empathizing with a patient motivated them to make correct decisions, even if this meant to ignore the DSS's advice in favor of their own choice, or vice versa. This required an understanding of when the DSS's advice would be correct and incorrect and how it would behave in novel situations.

The chapter is structured as follows. First we discuss the background and shortcomings of current XAI human evaluations. Furthermore, we provide examples on how *rule-based* and *example-based* explanations are currently used in XAI. The subsequent section describes three sets of recommendations for human evaluations in XAI, based on our experience designing the evaluation as well as on relevant literature. Next, we illustrate our own recommendations by explaining the use case in more detail and offering the theory behind our hypotheses. This is followed by a detailed description of our methods, analysis and results. We conclude with a discussion on the validity and reliability of the results and a brief discussion of future work.

2.2 Background

The following two sections discuss the current state of human evaluations in XAI and *rule-based* and *example-based* contrastive explanations. The former section illustrates the shortcomings of current evaluations, formed by either a lack of validity and reliability or the entire omission of an evaluation. The latter discusses the two explanation forms used in our evaluation in more detail, and illustrates their prevalence in the field of XAI.

2.2.1 Human evaluations in XAI

A major goal of Explainable Artificial Intelligence (XAI) is to have AI-systems construct explanations for their own output. Common purposes of these explanations are to increase system understanding [78], improve behaviour predictability [79] and calibrate system trust [80, 81, 74]. Other purposes include support in system debugging [82, 78], verification [79] and justification [83]. Currently, the exact purpose of explanation methods is often not defined or formalized, even though these different purposes may result in profoundly different requirements for explanations [32]. This makes it difficult for the field of XAI to progress and to evaluate developed methods.

The difficulties for human evaluations in XAI are reflected in recent surveys from Anjomshoae et al. [73], Adadi et al. [38], and Doshi-Velez and Kim [33] that summarize current efforts of evaluations in the field. The systematic literature review by Anjomshoae et al. [73] shows that 97% of the 62 reviewed articles underline that explanations serve a human need, 41% did not evaluate their explanations with humans. In addition, of those papers that performed a human evaluation, relatively few provided a good discussion of the context (27%), results (19%) and limitations (14%) of their experiment. The second survey from Adadi et al. [38] evaluated 381 papers and found that only 5% had an explicit focus on the evaluation of the XAI methods. These two surveys show that, the few evaluations provide limited conclusions which make it difficult for the field of XAI to progress based on these results.

A third survey by Doshi-Velez and Kim [33] discusses an explicit issue with human evaluations in XAI. The authors argue to systematically start evaluating different explanations classes and forms in various domains, a rigor that is currently lacking in the evaluations. To do so in a valid way, several recommendations are given. First, the application level of the study context should be made clear; either a real, simplified or generic application. Second, any (expected) task-specific explanation requirements should be mentioned. Examples include the average human level of expertise targeted, and whether the explanation should address the entire system or a single output. Finally, the explanations and their effects should be clearly stated together with a discussion of the study's limitations. Together, these three surveys illustrate the shortcomings of current human evaluations for XAI.

From several studies that do focus on evaluating the effects of explanations on humans, we note that the majority focuses on subjective measurements. Surveys and interviews are used to measure satisfaction [84, 85], the goodness of an explanation [86], acceptance of the system's advice [87, 88] and trust in the system [89, 90, 91, 92]. Such subjective measurements can provide a valuable insight in the human's perspective on the explanation. However, these results do not necessarily relate to the behavioural effects an explanation could cause. Therefore, these subjective measurements require further investigation to see if they correlate with a behavioural effect [61]. Without such an investigation, these subjective results only provide information on the human's beliefs and opinions, but not on actual gained understanding, trust or task performance. Some studies do perform objective measurements. The work from [93] for example, measured both subjective ease-of-use of an explanation and a participant's capacity to correctly make inferences based on the explanations. This allowed the authors to differentiate between behavioural and self-perceived effects of an explanation, underlining the value of performing objective measurements.

The above described critical view on human evaluations is related to the concepts of construct validity and that of reliability. These two concepts provide clear standards to scientifically sound human evaluations [94, 95, 96]. The construct validity of an evaluation is its accuracy in measuring the intended constructs (e.g. understanding or trust). Examples of how validity may be harmed is a poor design, ill-defined constructs or arbitrarily selected measurements. Reliability on the other hand refers to the evaluation's internal consistency and reproducibility, and may be harmed by a lack of documentation, an unsuitable use case or noisy measurements. In the social sciences, a common condition for results to be generalized to other cases and to infer causal relations is that a human evaluation is both valid and reliable [94]. This can be (partially) obtained by developing various types of measurements for common constructs. For example, self-reported subjective measurements such as ratings and surveys can be supplemented by behavioural measurements to gather data on the performance in a specific task.

2.2.2 Rule-based and example-based explanations

Human explanations tend to be contrastive: they compare a certain phenomenon (fact) with a hypothetical one (foil) [97, 98]. In the case of a decision support systems (DSS), a natural question to ask is 'Why this advice?'. This question implies a contrast, as the person asking this question often has an explicit contrasting foil in mind. In other words, the implicit question is 'Why this advice and not that advice?'. The specific contrast allows the explanation to be limited to the differences between fact and foil. Humans use contrastive explanations to explain events in a concise and specific manner [68]. This advantage also applies to systems: contrastive explanations narrow down the available information to a concrete difference between two outputs.

Contrastive explanations can vary depending on the way the advice is contrasted with a different advice, for example using rules or examples. Within the context of a DSS advising an insulin dose for DMT1 self-management, a contrastive rule-based explanation could be: "Currently the temperature is below 10 degrees and a lower insulin dose is advised. If the temperature was above 30 degrees, a normal insulin dose would have been advised." This explanation contains two rules that explicitly state the differentiating decision boundaries between the fact and foil. Several XAI methods aim to generate these "if ... then ..." rules [99, 100, 57, 101].

An example-based explanation refers to historical situations in which the advice was found to be true or false: "The temperature is currently 8 degrees, and a lower insulin dose is advised. Yesterday was similar: it was 7 degrees and the same advice proved to be correct. Two months ago, when it was 31 degrees, a normal dose was advised instead, which proved to be correct for that situation". Such example- or instance-based explanations are often used between humans, as they illustrate past behaviour and allow for generalisation to new situations [102, 103, 104, 105]. Several XAI methods try to identify examples to generate such explanations, for example those from [106, 107, 108, 109, 110].

Research on system explanations using rules and examples is not new. Most of this research focused on exploring how humans preferred a system would reason, by rules or through examples. For example, humans prefer an example-based spam-filter over a rule-based [111], while they prefer spam-filter explanations to be rule-based [112]. Another evaluation showed that the number of rule factors in an explanation had an effect on task performance by either promoting system over-reliance (too many factors) or self-reliance (too few factors) [113].

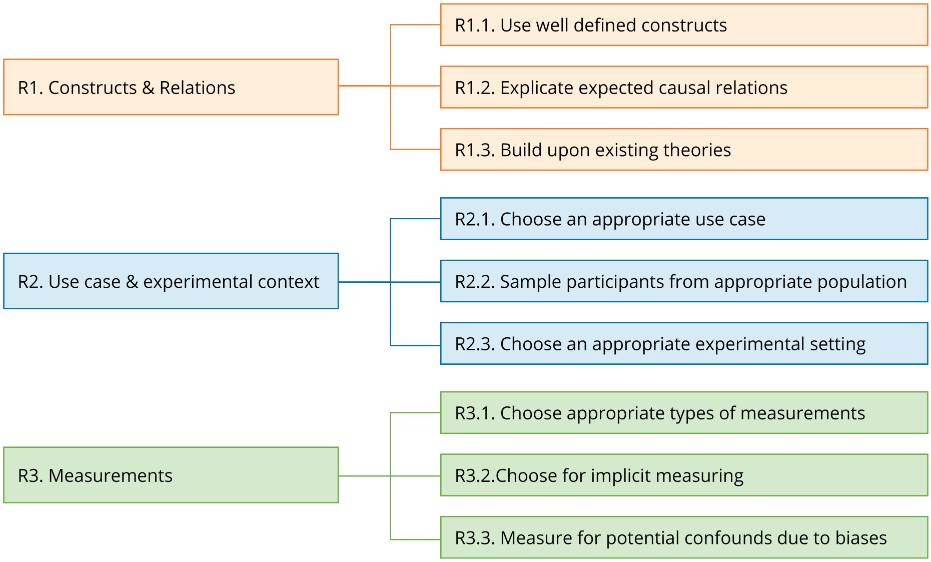

R1. Constructs & Relations
R1.1. Use well defined constructs
R1.2. Explicate expected causal relations
R1.3. Build upon existing theories
R2. Use case & experimental context
R2.1. Choose an appropriate use case
R2.2. Sample participants from appropriate population
R2.3. Choose an appropriate experimental setting
R3. Measurements
R3.1. Choose appropriate types of measurements
R3.2.Choose for implicit measuring
R3.3. Measure for potential confounds due to biases

Figure 2.1: An overview of three sets of practical recommendations to improve human evaluations in XAI.

Work by Lim et al. [114] shows that rule-based explanations cause humans to understand system behaviour, especially if those rules explain why the system behaves in a certain way as opposed to why it does not behave in a different (expected) way. Studies such as these tend to evaluate either rules or examples, depending on the research field (e.g. recommender system explanations tend to be example-based) but few compare rules with examples.

2.3 Recommendations for human evaluations in XAI

As discussed in Section 2.2.1, human evaluations play an invaluable role in XAI but are often omitted or of insufficient quality. Our main contribution is a thorough evaluation of rule-based and example-based contrastive explanations. In addition, we believe that the experience and lessons learned in designing this evaluation can be valuable for other researchers. Especially researchers in the field of XAI that are less familiar with human evaluations can benefit from guidance in their design available from other research communities. To that end, we propose three sets of recommendations with practical methods to help improve human evaluations in XAI. An overview is provided in Figure 2.1.

2.3.1 R1: Constructs and relations

As stated in Section 2.2.1, the field of XAI often deals with ambiguously defined concepts such as ‘understanding’. We believe that this hinders the creation and replication of evaluations and their results. Through clear definitions and motivation, the contribution of the evaluation becomes more apparent. This also aids other researchers to extend on the results. We provide three practical recommendations to clarify the evaluated constructs and their relations.

Our first recommendation is to clearly define the intended purposes of an explanation in the form of a construct. A construct is either the intended purpose, an intermediate requirement for the purpose or a potential confound to your purpose. Constructs form the basis of the scientific theory underlying XAI methods and human evaluations. By defining a construct it becomes easier to develop measurements. Second, we recommend to clearly define the relations expected between the constructs. A concrete and visual way to do so is through a Causal Diagram which presents the expected causal relations between constructs [115]. These relations form your hypotheses and make sure they are formulated in terms of your constructs. Clearly stating hypotheses allow other researchers to critically reflect on the underlying theory assumed, proved or falsified with the evaluation. It offers insight in how constructs are assumed to relate and how the results support or contradict these relations.

Our final recommendation regarding constructs is to adopt existing theories, such as from philosophy, (cognitive) psychology and from human-computer interaction (see [68, 42] for an overview). The former provides construct definitions whereas the latter two provide theories of human-human and human-computer explanations. These three recommendations to define constructs and their relations and grounding them in other research disciplines can contribute to more valid and reliable human evaluations. In addition, this practice allows results to be meaningful even if hypotheses are rejected, as they falsify a scientific theory that may have been accepted as true.

2.3.2 R2: Use case and experimental context

The second set of recommendations regards the experimental context, including the use case. The use case determines the task, participants that can and should be used, the mode of the interaction, the communication that takes place and the information available to the human [116]. As Doshi-Velez and Kim already stated, the selected use case has a large effect on the conclusions that can be drawn and the extent to which they can be generalized [33]. Also, the use case does not necessarily need to be of high fidelity, as a low fidelity allows for more experimental control and a potentially more valid and reliable evaluation [117]. We recommend to take these aspects into account when determining the use case and to reflect on the choices made when interpreting the results from the evaluation. This improves both the validity and reliability of the evaluation. A concrete way to structure the choice for a use case is to follow the taxonomy provided by Doshi-Velez and Kim [33] (see Section 2.2.1) or a similar one.

The second recommendation concerns the sample of participants selected, as this choice determines the initial knowledge, experience, beliefs, opinions and biases humans have. Whether participants are university students, domain experts or recruited online through platforms such as Mechanical Turk, the characteristics of the group will have an effect on the results. The choice of population should be governed by purpose of the evaluation. For example, our evaluation was performed with healthy participants rather than diabetes patients as those tend to vary in their diabetes knowledge and suffer from misconceptions [118]. These factors can interfere in an exploratory study such as ours, whose findings are not domain specific. Hence, we recommend to invest in both understanding the use case domain and reflecting on the intended purpose of the evaluation. These considerations should be consolidated in inclusion criteria to ensure that the results are meaningful with respect to the study's aim.

Our final recommendation related to the context considers the experimental setting and surroundings, as these may affect the quality and generalizability of the results. An online setting may provide a large quantity of readily available participants, but the results are often of ambiguous quality (see Paolacci et. al for a review [119]). If circumstances allow, we recommend to use a controlled setting (e.g. a room with no distractions, or a use case specific environment). This allows for valuable interaction with participants while reducing potential confounds that threaten the evaluation's reliability and validity.

2.3.3 R3: Measurements

Multiple measurements exist for computational experiments on suggested XAI methods (for example; fidelity [120], sensitivity [121] and consistency [122]). However, there is a lack of validated measurements for human evaluations [61]. Hence, our third group of recommendations regards the type of measurement to use for the operationalization of the constructs. We identify two main measurement types useful for evaluations: self-reported measures and behavioural measures. Self-reported measures are subjective and are often used in evaluations in XAI. They provide insights in a human's conscious thoughts, opinions and perceptions. We recommend the use of self-reported measures for subjective constructs (e.g. perceived understanding), but also recommend a critical perspective on whether the measures indeed address the intended constructs. behavioural measures have a more observational nature and are used to measure actual behavioural effects. We recommend their usage for objectively measuring constructs such as understanding and task performance. Importantly however, such measures often only measure one aspect of behaviour. Ideally, a combination of both measurement types should be used to assess effects on both the human's perception and behaviour. In this way, a complete perspective on a construct can be obtained. In practice, some constructs lend themselves more for self-reported measurements, for example a human's perception on trust or understanding. Other constructs are more suitable for behavioural measurements, such as task performance, simulatability, predictability, and persuasive power.

Furthermore, we recommend to measure explanation effects implicitly rather than explicitly. When participants are not aware of the evaluation's purpose, their responses may be more genuine. Also, when measuring understanding or similar constructs, the participant's explicit focus on the explanations may cause skewed results not present in a real world application. This leads to our third recommendation to measure potential biases. Biases can regard the participant's overall perspective on AI, the use case, decision making or similar. However, biases can also be introduced by the researchers themselves. For example, one XAI method can be presented more attractively or reliably than another. It can be difficult to prevent such biases. One way to mitigate these biases is to design how the explanation are presented, the explanation form, in an iterative manner with expert reviews and pilots. In addition, one can measure these biases nonetheless if possible and reasonable. For example, a usability questionnaire can be used to measure potential differences between the way explanations are presented in the different conditions. For our study we designed the explanations iteratively and verified that the chosen form for each explanation class did not differ significantly in the perception of the participants.

2.4 The use case: diabetes self-management

In this study, we focused on personalized healthcare, an area in which machine learning is promising and explanations are essential for realistic applications [123]. Our use case is that of assisting patients with diabetes mellitus type 1 (DMT1) with personalized insulin advice. DMT1 is a chronic autoimmune dis-order in which glucose homeostasis is disturbed and intake of the hormone insulin is required to balance glucose levels. Since blood glucose levels are influenced by both environmental and personal factors, it is often difficult to find the adequate dose of insulin that stabilizes blood glucose levels [124]. Therefore, personalized advice systems can be a promising tool in DMT1 management to improve quality of life and mitigate long-term health risks.

In our context, a DMT1 patient finds it difficult to find the optimal insulin dose for a meal given a situation. On the patient's request, a fictitious intelligent DSS provides assistance with the insulin intake before a meal. Based on different internal and external factors (e.g. hours of sleep, temperature, past activity, etc.), the system may advise to take a higher, lower or normal insulin dose. For example, the system could advise a lower insulin dose based on the current temperature. The factors that were used in the evaluation are realistic, and were based on Bosch [125] and an interview with a DMT1 patient.

In this use case, both the advice and the explanations are simplified. This study therefore falls under the human grounded evaluation category of Doshi-Velez and Kim [33]: a simplified task of a real-world application. The advice is binary (higher or lower), whereas in reality one would expect either a specific dose or a range of suggested doses. This simplification allowed us to evaluate with novices (see Section 2.6.3), as we could limit our explanation to the effects of a too low or too high dosage without going into detail about effects of specific doses. Furthermore, this prevented the unnecessary complication of having multiple potential foils for our contrastive explanations. Although the selection of the foil, either by system or human, is an interesting topic regarding contrastive explanations, it was deemed out of scope for this evaluation. The second simplification was that the explanations were not generated using a specific XAI method, but designed by the researchers instead. Several design iterations were conducted based on feedback from XAI researchers and interaction designers to remove potential design choices in the explanation form that could cause one explanation to be favored over another. Since the explanations were not generated by a specific XAI method, we were able to explore the effects of more prototypical rule- and example-based explanations inspired by multiple XAI methods that generate similar explanations (see Section 2.2.2).

There are several limitations caused by these two simplifications. First, we imply that the system can automatically select the appropriate foil for contrastive explanations. Second, we assume that the XAI method is able to identify only the most relevant factors to explain a decision. Although this assumes a potentially complex requirement for the XAI method, it is a reasonable assumption as humans prefer a selective explanation over a complete one [68].

2.5 Constructs, expected relations and measurements

Our evaluation focused on three constructs: system understanding, persuasive power, and task performance. Although an important goal of offering explanations is to allow humans to arrive at the appropriate level of trust in the system [126, 61], the construct of trust is difficult to define and measure [32]. As such, our focus was on constructs influencing trust that were more suitable to translate into measurable constructs; the intermediate construct of system understanding and the final construct of task performance of the entire human-system collaboration. The persuasive power of an explanation was also measured, as an explanation might cause a human to overly trust the system; believing that the system is correct while it is not, without having a proper system understanding. As such, the persuasive power of an explanation confounds to the effect of understanding on task performance.

Both *contrastive rule-* and *example-based* explanations were compared to each other with *no explanation* as a control. Our hypotheses are visualized in Figure 2.2, as a Causal Diagram [115]. From *rule-based* explanations we expected participants to gain a better understanding of when and how the system arrives at a specific advice. *Contrastive rule-based* explanations explicate the system's decision boundary between fact and foil and we expected the participants to recall and apply this information. Second, we expected that *contrastive example-based* explanations persuade participants to follow the advice more often. We believe that examples raise confidence in the correctness of an advice as they illustrate past good performance of the system. Third, we hypothesized that both system understanding and persuasive power have an effect on task performance. Whereas this effect was expected to be positive for system understanding, persuasive power was expected to affect task performance negatively in case a system's advice is not always correct. This follows the argumentation that persuasive explanations can cause harm as they may convince humans to over-trust a system [58]. Note that we conducted two separate experiments to measure the effects of an explanation class on understanding and persuasion. This allowed us to measure the effect of each construct separately on task performance, but not their combined effect (e.g. whether sufficient understanding can counteract the persuasiveness of an explanation).

The construct of understanding was measured with two behavioural measurements and one self-reported measurement. The first behavioural measure-

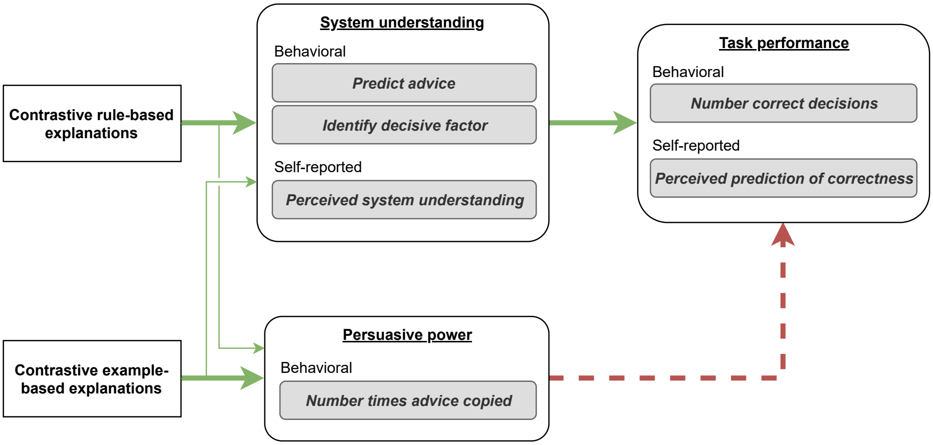


Figure 2.2: Our theory, depicted as a Causal Diagram. It describes the expected effects of contrastive rule- and example-based explanations on the constructs of system understanding, persuasive power and task performance. The solid green arrows depict expected positive effects and the red dashed arrow depicts a negative effect. The arrow thickness depicts the size of the expected effect. The opaque grey boxes are the measurements we performed for that construct, divided in either behavioural or self-reported measurements.

ment assessed the participant's capacity to correctly *identify the decisive factor* of the situations in the system's advice. This measured to what extent the participant recalled what factor the system believed to be important for a specific advice and situation. Second, we measured the participant's ability to accurately *predict the advice* in novel situations. This tested whether the participant obtained a mental model of the system that was sufficiently accurate enough to predict its behaviour in novel situations. The self-reported measurement tested the participant's *perceived system understanding*. This provided insight in whether participants over- or underestimated their understanding of the system compared to what their behaviour told us.

Persuasive power of the system's advice was measured with one behavioural measurement, namely the number of times participants *copied the advice*, independent of its correctness. If participants followed the advice with an explanation more often than participants without an explanation, we addressed this to the persuasiveness of the explanation.

Task performance was measured as the *number of correct decisions*, a behavioural measurement, and *perception of predicting advice correctness*, a self-reported measurement. We assumed a system that did not have a 100% accurate performance, meaning that it also made incorrect decisions. Therefore, the number of correct decisions made by the participant while aided by the system could be used to measure task performance. The self-reported measure allowed us to measure how well participants believed they could predict the correctness of the system advice.

Finally, two self-reported measurements were added to check for potential confounds. The first was a brief *usability questionnaire* addressing issues such as readability and the organisation of information. This could reveal whether one explanation form was designed and visualized better than the other, which would be a confounding variable. The second, *perceived system accuracy*, measured how accurate the participant thought the system was. This could help identify a potential over- or underestimation of the usefulness of the system, that could have affected to what extent participants attended to the system's advice and explanation.

The combination of self-reported and behavioural measurements enabled us to draw relations between our observations and a participant's own perception. Finally, by measuring a single construct with different measurements (known as triangulation [127]) we could identify and potentially overcome biases and other weaknesses in our measurements.

2.6 Methods

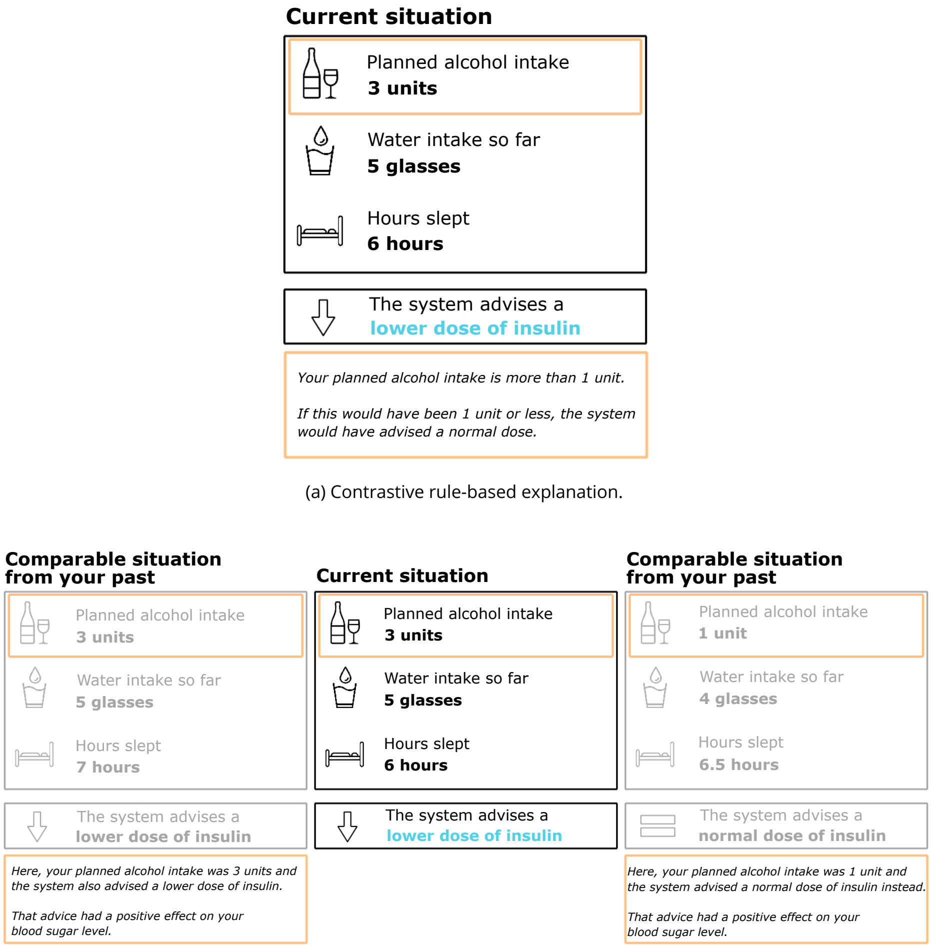
In this section we describe the operationalization of our evaluation in two separate experiments in the context of DSS advice in DMT1 self-management (see Section 2.4). *Experiment I* focused on the construct of system understanding. *Experiment II* focused on the constructs persuasive power and task performance. The explanation form (contrastive rule-based, contrastive example-based or no explanation) was the independent variable in both experiments and was tested between-subjects. See Figure 2.3 for an example of each explanation form.

The experimental procedure was similar in both experiments:

1. **Introduction**. Participants were informed about the study, use-case and task, and presented with a brief narrative about a DMT1 patient for immersion.

2. **Demographics questionnaire**. Age and education level were inquired to verify a sufficiently broad population sample.

3. **Pre-questionnaire**. Participants were questioned on DMT1 knowledge to assess if DMT1 was sufficiently introduced and to verify that participants lacked additional domain knowledge.



Current situation

Planned alcohol intake
3 units

Water intake so far
5 glasses

Hours slept
6 hours

The system advises a
lower dose of insulin

Your planned alcohol intake is more than 1 unit.

If this would have been 1 unit or less, the system would have advised a normal dose.

(a) Contrastive rule-based explanation.

Comparable situation from your past

Planned alcohol intake
3 units

Water intake so far
5 glasses

Hours slept
7 hours

The system advises a
lower dose of insulin

Here, your planned alcohol intake was 3 units and the system also advised a lower dose of insulin.

That advice had a positive effect on your blood sugar level.

Current situation

Planned alcohol intake
3 units

Water intake so far
5 glasses

Hours slept
6 hours

The system advises a
lower dose of insulin

Comparable situation from your past

Planned alcohol intake
1 unit

Water intake so far
4 glasses

Hours slept
6.5 hours

The system advises a
normal dose of insulin

Here, your planned alcohol intake was 1 unit and the system advised a normal dose of insulin instead.

That advice had a positive effect on your blood sugar level.

(b) Contrastive example-based explanation.

Figure 2.3: The two explanation forms. Both explanations were contrastive. Participants could view the situation, advice and explanation indefinitely.

4. **Learning block**. Multiple stimuli were presented with either the example- or rule-based explanations, or no explanations (control group).

5. **Testing block**. Several trials to conduct the behavioural measurements (*advice prediction* and *decisive factor identification* in *Experiment I*, the *number of advice copied* and *number of correct decisions* in *Experiment II*).

6. **Post-questionnaire**. A questionnaire to obtain the measurements (*perceived system understanding* in *Experiment I* and *perceived prediction of advice correctness* in *Experiment II*).

Factor	Insulin dose	Exp. I Rules	Exp. II Rules
Planned alcohol intake	Lower dose	> 1 unit	> 1 unit
Planned physical exercise	Lower dose	> 17 minutes	> 20 minutes
Physical health	Lower dose	Diarrhoea & Nausea	Diarrhoea & Nausea
Hours slept	Higher dose	< 6 hours	< 6 hours
Environmental temperature	Higher dose	$>26\,°\mathrm{C}$	$>31\,°\mathrm{C}$
Anticipated tension level	Higher dose	> 3 (a little tense)	> 4 (quite tense)
Water intake so far	-	-	-
Planned caffeine intake	-	-	-
Mood	-	-	-

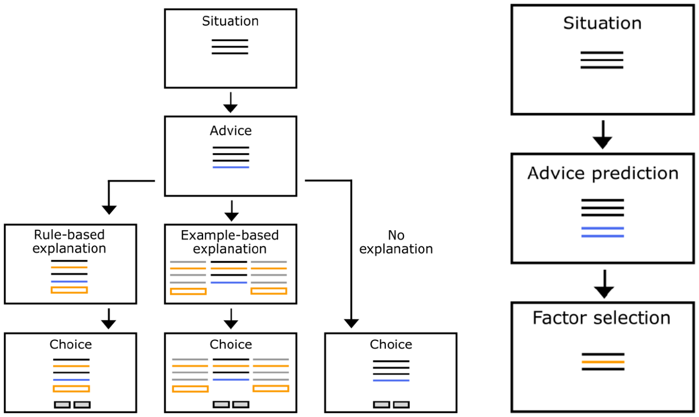
Table 2.1: An overview of the nine factors that played a role in the experiment. For each factor its influence on the true insulin dose is shown and the system threshold for that influence. These differed between the two experiments, the set of rules of the first experiment were defined as the ground truth. Three factors acted as fillers and had no influence.

7. **Usability questionnaire**. Participants filled out a usability questionnaire to identify potential interface related confounds.

8. **Control questionnaire**. The experimental procedure concluded with several questions to assess whether the purpose of the study was suspected and to measure *perceived system accuracy* to identify over- or under-trust in the system.

2.6.1 Experiment I: System understanding

The purpose of *Experiment I* was to measure the effects of *rule-based* and *example-based* explanations on system understanding compared to each other and to the control group with *no explanations*. See Figure 2.4 for an overview of both the learning and testing blocks. The learning block consisted of 18 randomly ordered trials, each trial describing a single situation with three factors and values from



(a) Learning block, Experiment I.

(b) Testing block, Experiment I.

Figure 2.4: A schematic overview of the learning and testing blocks of Experiment I.

Table 2.1. The situation description was followed by the system's advice, in turn followed by an explanation (in the experimental groups). Finally, the participant was asked to make a decision on administering a higher/lower insulin dose. This block served only to familiarize the participant to the system's advice and its explanation and to learn when and why a certain advice was given. Participants were not instructed to focus on the explanations in the learning block, nor were they informed of the purpose of the two blocks.

In the testing block, two behavioural measures were used to test the construct of understanding: *predicted advice* and *decisive factor identification*. The testing block consisted of 30 randomized trials, each with a novel situation description. Each description was followed by the question what advice the participant thought the system would give. This formed the measurement of *advice prediction*. The measurement *decisive factor identification* was formed by the subsequent question to select a single factor from a situation description that they believed was decisive for the predicted system advice.

A third, self-reported measurement was conducted in the post-questionnaire, which contained an eight-item questionnaire based on a 7-point Likert scale. These items formed the measurement of *perceived gained understanding*. The questions were asked without mentioning the term explanation and simply addressed 'system output'. The eight items were deemed necessary, to obtain a measurement less dependent on the formulation of one item.

2.6.2 Experiment II: Persuasive power and task performance

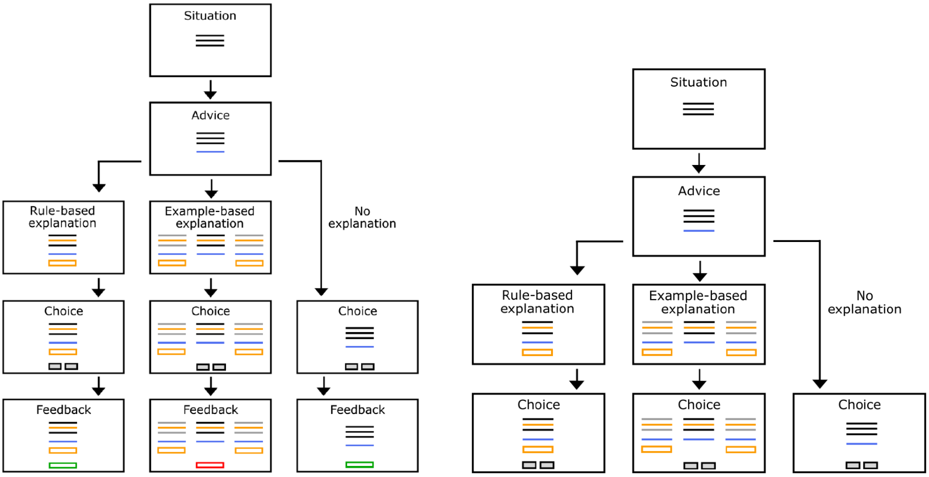
The purpose of *Experiment II* was to measure the effects of *rule-based* and *example-based* explanations on persuasive power and task performance, and to compare these to each other and to the control group with *no explanation*. Figure 2.5 provides an overview of the learning and testing blocks of this experiment. The learning block was similar to that of the first experiment: a situation was shown, containing three factors from Table 2.1. In the experimental groups, the situation was followed by an advice and explanation. Next, the participant was asked to make a decision on the insulin dose. After this point, the learning block differed from the learning block in the first experiment: the participant's decision was followed with feedback on its correctness. In 12 of the 18 randomly ordered trials of this learning block (66%), the system's advice was correct. In the six other trials, the advice was incorrect. Through this feedback, participants learned that the system's advice could be incorrect and in which situations. Instead of following the ground truth rule set (from *Experiment I)*, this system followed a second, partially correct set of rules, as shown in Table 2.1.

The testing block contained 30 trials, also presented in random order, in which a presented situation was followed by the system's advice and a potential explanation. Next, participants had to choose which insulin dose was correct based on the system's advice, explanation and gained knowledge of when the system is incorrect. Persuasive power was operationalized as the number of times a participant followed the advice, independent of whether it was correct or not. Task performance was represented by the number of times a correct decision was made. The former reflected how persuasive the advice and explanation was, even when participants experienced system errors. The latter reflected how well participants were able to understand when the system makes errors and compensate accordingly in their decision.

Also in this experiment, a self-reported measurement with eight 7-point Likert scale questions was performed. It measured the participant's subjective sense of their ability to estimate when the system was correct.

2.6.3 Participants

In *Experiment I*, 45 participants took part, of which 21 female and 24 male, aged between 18 and 64 years old ($M = 44.2 \pm 16.8$). Their education levels varied from lower vocational to university education. In *Experiment II* 45 different participants took part, of which 31 females and 14 males, aged between 18 and 61 years old ($M = 36.5 \pm 14.5$). Their education levels varied from secondary vo-



(a) Learning block, Experiment II.

(b) Testing block, Experiment II.

Figure 2.5: A schematic overview of the learning and testing blocks of Experiment II.

cational to university education. Participants were recruited from a participant database at TNO Soesterberg (NL) as well as via advertisements in Utrecht University (NL) buildings and on social media. Participants received a compensation of 20,- euro and their travel costs were reimbursed. Both samples represented the entire Dutch population and as such the entire range of potential DMT1 patients, hence the wide age and educational ranges.

The inclusion criteria were as follows: not diabetic, no friends or close relatives with diabetes, and no extensive knowledge of diabetes through work or education. General criteria were Dutch native speaking, good or corrected eyesight, and basic experience using computers. These inclusion criteria were verified in the pre-questionnaire. A total of 16 participants reported a close relative or friend with diabetes and one participant had experience with diabetes through work, despite clear inclusion instructions beforehand. After careful inspection of their answers, none were excluded because their answers on diabetes questions in the pre-questionnaire were not more accurate or elaborate than others. From this we concluded that their knowledge of diabetes was unlikely to influence the results.

2.7 Data analysis

Statistical tests were conducted using SPSS Statistics 22. An alpha level of 0.05 was used for all statistical tests.

The data from the measures in *Experiment I* were analyzed using a one-way Multivariate Analysis of Variance (MANOVA) with explanation form (*rule-based*, *example-based* or *no explanation*) as the independent between-subjects variable and *predicted advice* and *identified decisive factor* as dependent variables. The reason for a one-way MANOVA was due to the multivariate operationalization of a single construct, understanding [128]. Cronbach's Alpha was used to assess the internal consistency of the self-reported measurement for *perceived system understanding* from the post-questionnaire. Subsequently, a one-way Analysis of Variance (ANOVA) was conducted with the mean rating on this questionnaire as dependent variable and the explanation form as independent variable. Finally, the relation between the two behavioural and the self-reported measurements was examined with Pearson's product-moment correlations.

For *Experiment II* two one-way ANOVA's were performed. The first ANOVA had the explanation form (*rule-based*, *example-based* or *no explanation*) as independent variable and the *number of times the advice was copied* as dependent variable. The second ANOVA also had explanation form as independent variable, but *number of correct decisions* as dependent variable. The internal consistency of the self-reported measurement of *perceived prediction of advice correctness* from the post-questionnaire was assessed with Cronbach's Alpha and analyzed with a one-way ANOVA. Explanation form was the independent and the mean rating on the questionnaire the dependent variable. The presence of correlations between the behavioural and the self-reported measurements was assessed with Pearson's product-moment correlations. Detected outliers were excluded from the analysis.

2.8 Results

2.8.1 Experiment I: System understanding

The purpose of *Experiment I* was to measure gained system understanding when a system provides a rule- or example-based explanation, compared to no explanation. This was measured with two behavioural measures and one self-reported measure.

Figure 2.6 shows the results on the two behavioural measures: correct advice

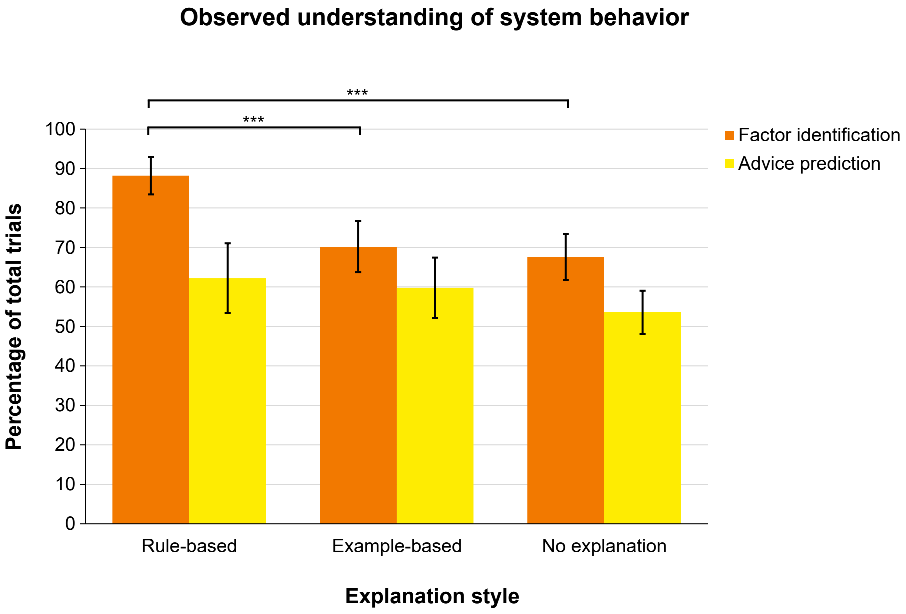


Figure 2.6: Bar plot of the mean percentages on correct prediction of the system's advice and correct identification of the decisive factor for that advice. Values are relative to the 30 randomized trials in Experiment I. The error bars represent a 95% confidence interval. *Note;* *** $p < 0.001$

prediction in novel situations and correct identification of the system's decisive factor. A one-way MANOVA with Wilks' lambda indicated a significant main effect of explanation form on both measurements ($F(4,82) = 6.675$, $p < 0.001$, $\Delta = .450$, $\eta_p^2 = .246$). Further analysis revealed a significant effect for explanation form on factor identification ($F(2,42) = 14.816$, $p < 0.001$, $\eta_p^2 = .414$), but not for advice prediction ($F(2,42) = 14.816$, $p = .264$, $\eta_p^2 = .414$). One assumption of a one-way MANOVA was violated, as the linear relationships between the two dependent variables and each explanation form was weak. This was indicated by Pearson's product-moment correlations for the rule-based $r = .487$ ($p = .066$), example-based $r = -.179$ ($p = .522$) and no explanation $r = .134$ ($p = .636$) groups. Some caution is needed in interpreting these results, as this lack of significant correlations shows a potential lack of statistical power. Further post-hoc analysis showed a significant difference in factor identification in favor of rule-based explanations compared to example-based explanations and no explanations ($p < 0.001$). No significant difference between example-based explanations and no explanation was found ($p = .796$).

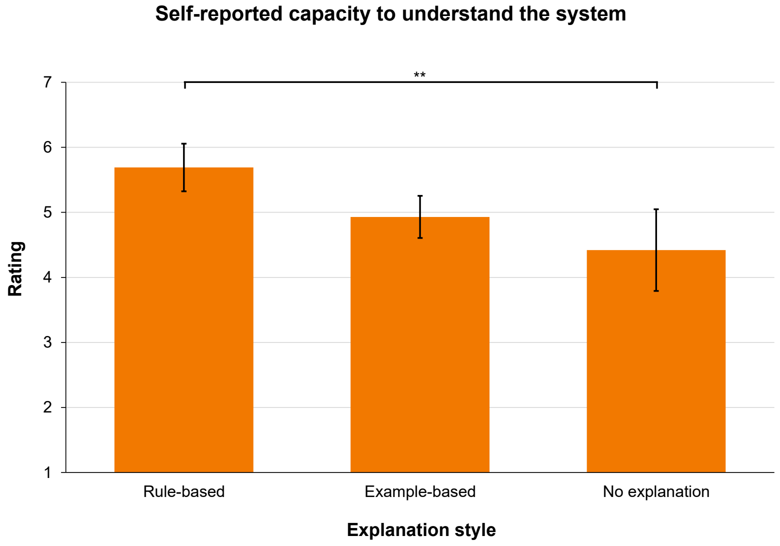


Figure 2.7: Bar plot of the mean self-reported system understanding. All values are on a 7-point Likert scale and error bars represent 95% confidence interval. *Note;* ** $p < 0.01$

Figure 2.7 shows the results on the self-reported measure of system understanding. The consistency between the different items in the measure was very high, as reflected by Cronbach's alpha ($\alpha = .904$). The mean rating over all eight items was used as the participant's subjective rating of system understanding. A one-way ANOVA showed a significant main effect of explanation form on this rating ($F(2,41) = 7.222$, $p = .002$, $p\eta_p^2 = .261$). Two assumptions of a one-way ANOVA were violated. First, the rule-based explanations group had one outlier, which did not affect the analysis in any way. Second, Levene's test was not significant ($p = .017$) signalling inequality between group variances. However, ANOVA is robust against the variance homogeneity violation with equal group sizes [129, 130]. Further post-hoc tests revealed that only rule-based explanations caused a significantly higher self-reported understanding compared to no explanations ($p = .001$). No significant difference was found for example-based explanations with no explanations ($p = .283$) and with rule-based explanations ($p = .072$).

Finally, Figure 2.8 shows a scatter plot between both behavioural measures and the self-reported measure. Pearson's product-moment analysis revealed no significant correlations between self-reported understanding and advice prediction ($r = -.007$, $p = .965$), not within the rule-based explanation group ($r = -.462$, $p = .129$), the example-based explanation group ($r = -.098$, $p = .729$), nor the no explanation group ($r = .001$, $p = .996$). Similar results were found

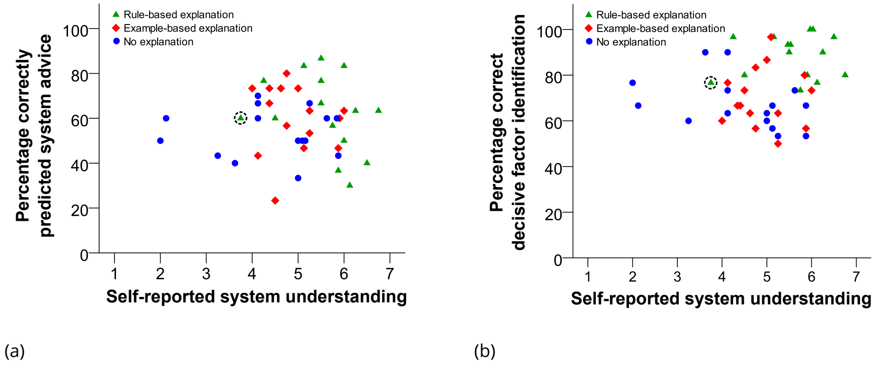


Figure 2.8: Scatter plots displaying the relation between a) advice prediction and b) decisive factor identification with self-reported understanding. Outliers are circled.

for the correlation between self-reported understanding and factor identification ($r = .192$, $p = .211$) and for the separate groups of rule-based explanations ($r = -.124$, $p = .673$), example-based explanations ($r = .057$, $p = .840$) and no explanations ($r = -.394$, $p = .146$).

2.8.2 Experiment II: Persuasive power and task performance

The purpose of *Experiment II* was to measure a participant's ability to use a decision support system appropriately when it provides a rule- or example-based explanation, compared with no explanation. This was measured with one behavioural and one self-reported measurement. In addition, we measured the persuasiveness of the system for each explanation form, compared to no explanations. This was assessed with one behavioural measure.

Figure 2.9 shows the results of the behavioural measure for task performance, as reflected by the huma's decision accuracy. A one-way ANOVA showed no significant differences ($F(2, 41) = 1.716$, $p = .192$, $\eta_p^2 = .077$). Two violations of ANOVA were discovered. There was one outlier in the example-based explanations, with 93.3% accuracy (1 error). Removal of the outlier did not affect the analysis. Levene's test showed there was no homogeneity of variances ($p = .007$), however ANOVA is believed to be robust against this under equal group sizes [129, 130].

Figure 2.9 shows the results of the behavioural measure for persuasiveness, i.e. the number times system advice was followed. Note that in *Experiment II* the system's accuracy was 66.7%. Thus, following the advice in a higher percentage of cases denotes an adverse amount of persuasion. A one-way ANOVA showed

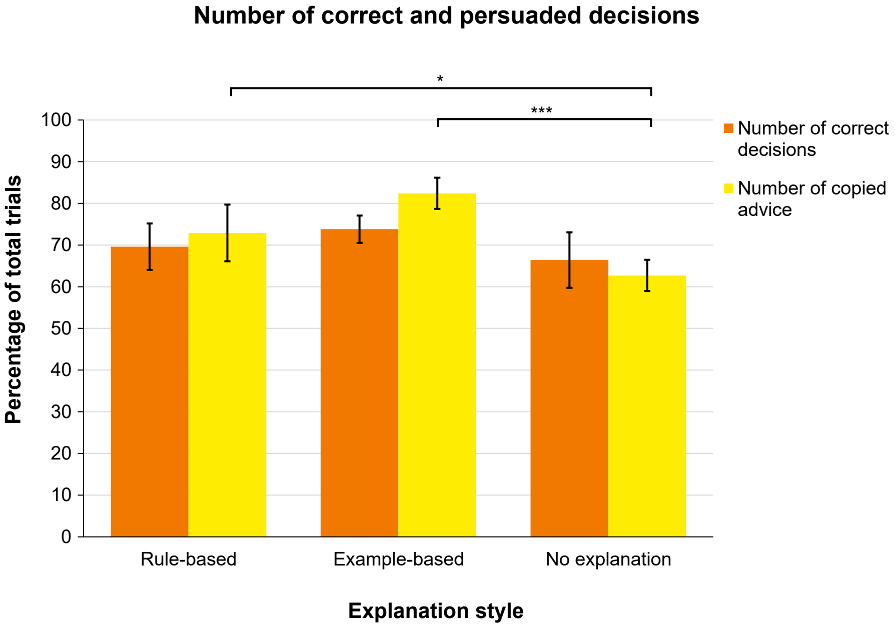


Figure 2.9: Results on task performance and persuasiveness as the mean percentage of correct decisions (a) and percentage of decisions similar to system's advice independent of correctness (b) respectively. Error bars represent a 95% confidence interval. *Note;* * $p < 0.05$, *** $p < 0.001$

that explanation form had a significant effect on following the system's advice ($F(2,41) = 11.593$, $p < .001$, $\eta_p^2 = .361$). Further analysis revealed that participants with no explanation followed the system's advice significantly less than those with rule-based ($p = .049$) and example-based explanations ($p < .001$). However, there was no significant difference between the two explanation forms ($p = .068$). One outlier violated the assumptions of an ANOVA. One participant in the rule-based explanation group followed the system's advice only 33.3% of the time. Its exclusion affected the outcomes of the ANOVA and the results after exclusion are reported.

Figure 2.10 displays the self-reported capacity to predict correctness, operationalized by a rating how well participants thought they were able to predict when system advice was correct or not. The consistency of the eight 7-point Likert scale questions was high according to Cronbach's Alpha ($\alpha = .820$). Hence, we took the mean rating of all questions as an estimate of participants' performance estimation. A one-way ANOVA was performed, revealing no significant differences ($F(2,41) = 2.848$, $p = .069$, $\eta_p^2 = .122$). One outlier from the rule-based

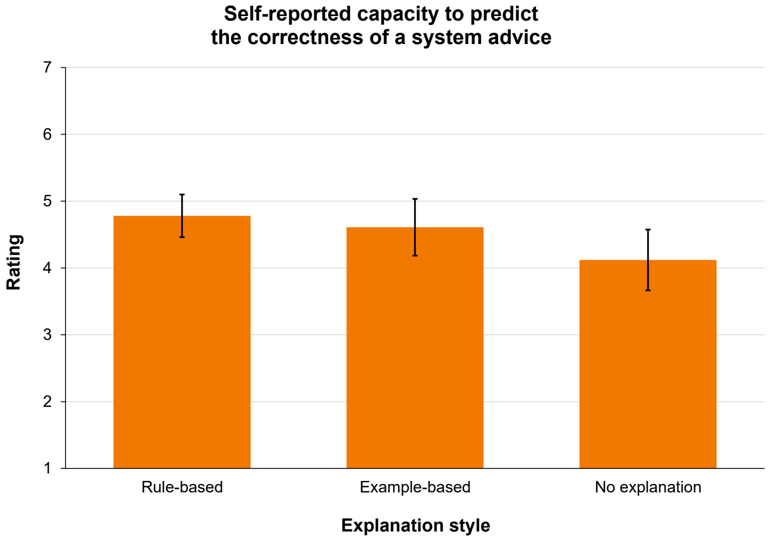


Figure 2.10: Bar plot of the mean self-reported system performance estimation. All values are on a 7-point Likert scale and error bars represent 95% confidence interval.

explanation group was found, its removal did not affect the analysis.

A correlation analysis was performed between the self-reported measurement of the predicted correctness and the behavioural measurement of making the correct decision, two measurements of task performance. The accompanying scatter plot is shown in Figure 2.11. A Pearson's product-moment correlation revealed no significant correlation between the self-reported and behavioural measure ($r = .146$, $p = .350$). Also, there were no significant correlations in the rule-based ($r = .411$, $p = .144$) and example-based explanation ($r = -.347$, $p = .225$) groups, or in the no explanation group ($r = .102$, $p = .718$). Both outliers from each measurement were removed in this analysis and did not affect the significance.

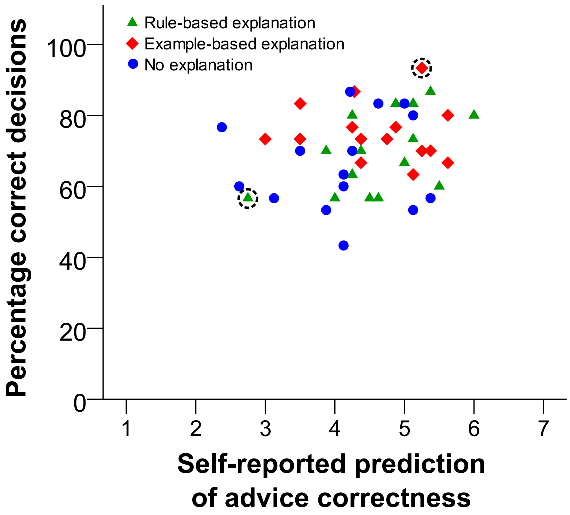


Figure 2.11: Scatter plot displaying the relation between number of correct decisions made and self-reported capacity to predict advice correctness. Outliers are circled.

2.8.3 Usability and biases

A usability questionnaire was used to evaluate if there were differences in usability between the two explanation forms, as this could influence the results. The questionnaire contained five questions on a 100-point scale about readability, organisation of information, language, images and color. The consistency between the five questions was relatively high, as revealed by a Cronbach's Alpha test ($\alpha = .722$). Figure 2.12 shows the mean ratings for each question, broken down by explanation form (*rule-based*, *example-based*, *no explanation*). No statistical analysis was performed, as this questionnaire only functioned as a check for potential usability confounds in the experiment.

In addition to the ratings, participants were asked about the positive and negative usability aspects of the system in two open questions. Common positive descriptions included “clear", “well-arranged", “clear and simple icons" and “understandable language". Although not many participants had negative remarks, most addressed insufficient visual contrast due to the colors used. Unique to the example-based explanations participant group were remarks about a lack of concise and well-arranged information.

In the control questionnaire we asked participants to give an estimate of the overall system's accuracy. This was to validate any potential overly positive or negative trust bias towards the system. In *Experiment I* the system was 100%

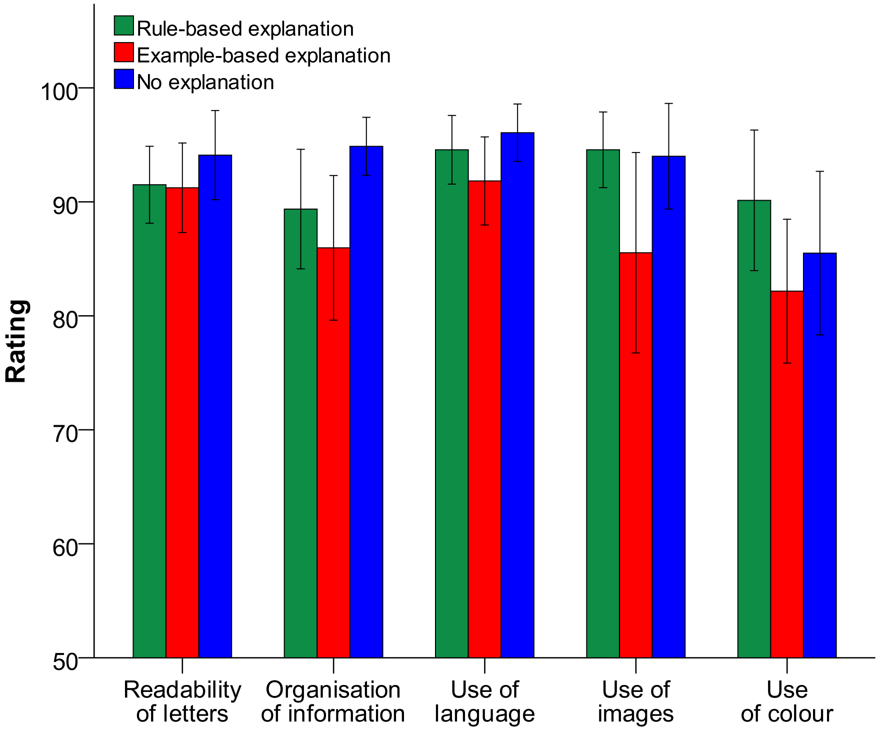


Figure 2.12: The mean ratings on the usability questions, separated on explanation form. The error-bars represent a 95% confidence interval.

accurate, but this was unknown to the participants since there was no feedback on correctness included. Nonetheless, estimates ranged from 30% to 90% ($\mu = 75.2\%$, $\sigma = 12.8\%$). This meant that all participants believed the system to make errors based on no information. In *Experiment II* the system's accuracy was 66.7%. Participants experienced this due to the feedback on made decisions in the learning block. Estimates ranged between 50% and 95% ($\mu = 74.8\%$, $\sigma = 8.8\%$), indicating that on average, system accuracy was overestimated.

After the experiment, brief discussions with participants revealed additional perspectives. Several participants from the no explanation group wished the system could give an explanation for its advice. One participant expressed a need for knowing the system's rules governing the system's advice. In the two explanation groups, participants experienced the explanations as useful. Rules were valued for there explicitness, whereas examples were viewed as inciting trust. However, in the two explanation groups several participants found it un-clear what the highlight of a factor (see Figure 2.3) meant. Several participants also mentioned that, although useful, the explanations lacked a causal rationale.

2.9 Discussion

Below we discuss the results from both experiments in detail and relate them to our theory presented in Section 2.5.

2.9.1 Experiment I: System understanding

Experiment I measured the participant's capacity to understand how and when the system provides a specific advice. This construct was operationalized in three measurements: *identification of decisive factor*, *predicting advice* and *perceived system understanding*. We hypothesized that participants receiving *contrastive rule-based explanations* would score best on all three measurements. *Contrastive example-based explanations* were only expected to improve understanding slightly more than *no-explanations* (see Figure 2.2).

The results from our evaluation support these hypotheses in part. First, *rule-based explanations* indeed seem to allow participants to more accurately identify the factor from a situation that was decisive in the system's advice. However, *rule-based* nor *example-based explanation* allowed participants to learn to predict system behaviour. The *rule-based explanations* however, did cause to participants to think that they better understood the system compared to *example-based* and *no explanations*. The *example-based explanations* only showed a small and insignificant increase in perceived system understanding. It is important to note that there was no correlation between the self-reported measurement of understanding and the behavioural measurements of understanding. This shows that participants had a perception of understanding that differed from the understanding as measured with *factor identification* and *advice prediction*.

Close inspection of the results showed two potential causes for the lack of support for our hypotheses. The first reason might be because the described DMT1 situations and accompanying system advice was too intuitive. This is supported by the fact that participants with *no explanation* were already quite adapt in *identifying decisive factors* (nearly 70% compared to 33% chance). The second reason we inferred from open discussions with participants after the experiment. Most participants who received either explanation form mentioned difficulty in applying and generalizing the knowledge from the explanations to novel situations. Several participants even expressed the desire to know the rationale of why a certain rule or behaviour occurred. This is in line with the theory that explanations should convey specific causal relations obtained from an overall causal model describing the behaviour of the system, instead of just factual correlations between system input and output.

If we generalize these results to the field of XAI, we have shown that *contrastive rule-based explanations* as "if ... then ..." statements are not sufficient to predict system behaviour. However, such explanations are capable of educating a human to identify which factors would play a decisive role in system advice given a specific situation. Also, such explanations seem to provide the human with the perception that (s)he is better capable of understanding the system. The *contrastive example-based explanations* however showed no improvement on observed or self-reported understanding. This experiment illustrated the need for explanations that provide more causal information, instead of solely information depicting system input and output correlations. Furthermore, we illustrated that self-reported and behavioural measurements of understanding may not correlate, underlining the need for (a combination of) measures that accurately and reliably measure the intended construct.

2.9.2 Experiment II: Persuasive power and task performance

In *Experiment II* we investigated the extent to which an explanation increases the persuasiveness of an advice, as well as the explanation's effect on task performance. The persuasive power of an explanation was operationalized with the *number of times the advice was copied*. Task performance was represented by the *number of correct decisions* and the self-reported *perception of predicting advice correctness*. We hypothesized that especially *contrastive example-based explanations* would increase persuasive power, while these in turn would lower actual task performance. In contrast, the understanding participants gained from *rule-based explanations* was expected to cause an increase in task performance (see Figure 2.2).

Both *contrastive rule-based* and *example-based* explanations showed more persuasive power than when *no explanation* was given. The *example-based explanations* also showed slightly more persuasive power than the *rule-based explanations*, but this difference was not significant. These results partly support our theory about persuasive power, as they illustrate that explanations persuade humans to follow a system's advice more often. These results however, do not support that *example-based explanations* are that much more persuasive than *rule-based explanations*.

With respect to task performance, we saw that explanations caused small but insignificant improvements on both behavioural and self-reported data. In fact, the *example-based explanations* showed the highest (but still insignificant) improvement. Due to a lack of statistical evidence not much can be inferred from this, and further evaluation is required.

Similar to *Experiment I* we found a lack of correlation between participants reporting their *perception of predicting advice correctness*, and the *number of correctly made decisions*. In other words, these measures do not seem to measure the same construct. An explanation could be that participants were unable to estimate their own capacity of predicting the correctness of advice.

We have shown that providing an explanation with an advice results in humans following that advice more often, even when incorrect. In addition, there was a suggestion that explanations also improve task performance, especially *contrastive example-based explanations*. However, these effects were marginal and not significant. These results underline the need in the field of XAI to take a different stance on which explanations should be generated. Two common forms of explanations answering a contrasting question did not appear to increase task performance, an effect often attributed to such explanations within the field.

2.10 Limitations

This study has several limitations that warrant caution in generalizing the results to other use cases or to the field of XAI in general. The first set of limitations is related to the selected use case of aided DMT1 self-management. This use case falls into the category 'simplified' from Doshi-Velez and Kim [33] as it approximates a realistic use case. However, two major aspects differ from the real-life situation. First, we recruited healthy participants who had to empathize with a DMT1 patient, instead of actual DMT1 patients. Nevertheless, participants were sampled from the entire Dutch population, resulting in a wide variety of ages and education. These choices allowed us to measure the effects of the explanation forms without focusing on a specific demographic or having to compensate for varying domain knowledge in DMT1 participants to correct for in the measured effects. Second, the system itself was fictitious and followed a predetermined set of rules rather than comprising the full complexity of a realistic system. These two simplifications prevent us to generalize the results and to apply our conclusions to construct an actual system for aiding DMT1 patients in self-management. However, this was not the purpose of this study. Instead, we aimed to evaluate whether the supposed effects of two often cited explanations forms were warranted. We believe the selected use case allowed us to do so, as it gave both context as well as motivation for the humans to understand explanations. Also, laymen were chosen opposed to DMT1 patients to mitigate any difference in diabetes knowledge and misconceptions which can vary greatly between patients (e.g. see Odili et. al [118]). Future research specifically targeted at the development of a DSS for DMT1 self-management should include DMT1 patients as participants.

The second set of limitations is related to suspected confounds in the experiment. A brief usability questionnaire showed that participants held an overall positive bias towards the system, whether an explanation was provided or not. In addition this questionnaire showed that participants' perception of the organisation of the information was not always positive. Hence, a potential limitation lies in the way the explanations were presented. Also, surprisingly, in *Experiment I* participants attributed a low performance to the system, while they had no information to do so. In *Experiment II* however, participants tended to slightly overestimate the system's actual performance. This occurred independent of the explanation form. This shows that the participants could have had a natural tendency to distrust the system's advice. This may have affected the self-reported results.

Finally, a few limitations arose from the design of both experiments. The results for the example-based explanations could have been different with a longer learning block, as it takes time to infer decision boundaries from examples. Also, both testing blocks were relatively long, which could have caused participants to continue learning about the system while we were measuring their understanding. We did not perform any analyses on this, as it would add another level of complexity to the design. Hence, we cannot say for certain that the learning block was of sufficient length to allow participants to learn enough from the explanations. However, if this was the case, we believe that prolonging the learning block would have only resulted in stronger effects. Lastly, due to the choice of different participant groups for both experiments, we could only draw limited conclusions on the relation between the understanding on the one hand and task performance and persuasiveness on the other hand. We selected this approach instead of combining the constructs in a single experiment with a within-subject design, to avoid learning effects not sufficiently compensated through randomizing the understanding and task performance/persuasion blocks.

2.11 Conclusion

A lack of human evaluations characterizes the field of Explainable Artificial Intelligence (XAI). A contribution of this paper was to provide a set of recommendations for future evaluations. Practical recommendations were given for XAI researchers unfamiliar with them. These addressed the evaluation's constructs and their relations, the selection of a use case and the experimental context, and suitable measurements to operationalize the constructs in the evaluation. These recommendations originated from our experience designing an extensive human evaluation. Our second contribution was to evaluate the effects of *contrastive rule-based* and *contrastive example-based explanations* on the participant's understanding of system behaviour, persuasive power of the system's advice when combined with an explanation, and task performance. The evaluation took place in a decision-support context where participants were aided in choosing the appropriate dose of insulin to mitigate the effects of diabetes mellitus type 1.

Results showed that *contrastive rule-based explanations* allowed participants to correctly identify the situational factor that played a decisive role in a system's advice. Neither *example-based* or *rule-based explanations* enabled participants to correctly predict the system's advice in novel situations, nor did they improve task performance. However, both explanation forms did cause participants to follow the system's advice more often, even when this advice was incorrect. This shows that both *rules* and *examples* that answer a contrastive question are not sufficient on their own to improve a human's understanding or task performance. We believe that the main reason for this is that these explanations lack a clarification of the underlying rationale of system behaviour.

Future work will focus on the evaluation of a combined explanation form provided in interactive form, to assess whether this interactive form helps humans to learn a system's underlying rationale. As an extension, potential methods will be researched that can generate causal reasoning traces, rather than decision boundaries, to expose the behaviour rationale directly. In addition, future work may focus on similar studies with actual diabetes patients to study potential homogeneous groups in terms of explanation effects (e.g. effect of age, domain knowledge, etc.). Finally, during the design and analysis of this evaluation we discovered a need for validated and reliable measurements. We will continue to use different types of measurements to measure constructs in a valid and reliable way in future evaluations.

2.12 Acknowledgements

We acknowledge the project ERP Hybrid Artificial Intelligence from TNO for funding this research. In addition, we thank the Technical University of Delft and the University of Amsterdam for support and feedback on this research.

CHAPTER 3

CONTRASTIVE EXPLANATIONS FOR CLASSIFICATION MODELS

This chapter is adapted from; **van der Waa, J.**, Robeer, M., van Diggelen, J., Brinkhuis, M., & Neerincx, M. (2018). Contrastive explanations with local foil trees. *Proceedings of the Workshop on Interpretable Machine Learning, International Joint Conference of Artificial Intelligence*. The adaptations include an altered abstract, shortened introduction, and an adjusted lay-out, including the formatting of figures and tables.

Jasper van der Waa, Marcel Robeer, Jurriaan van Diggelen, Matthieu Brinkhuis, Mark Neerincx

Explanation generating methods tend to aim for feature attributions; quantifying the relative importance of a feature in an AI's decision. These explanations tend to be superficial, as they do not explain what the feature's role is in that decision. In addition, the amount of information in the explanation grows with the number of referred to features, risking information overload and communicating redundant information. The previous chapter argued that contrastive explanations offer a natural way to limit the explained information to what is important by contrasting the current decision with a decision of interest. As the previous chapter illustrated the potential benefit of explaining this information in a rule-based form, this chapter proposes a method to generate rule-based contrastive explanations for an AI providing decision support. This method is based on the idea of training localized one-versus-all decision trees. These trees can then be used to identify the disjointed set of rules that causes an AI to make one decision instead of another within that area of the feature space. A technical evaluation of this method on three benchmark classification tasks reveals its efficiency and faithfulness to the AI's decision making.

3.1 Introduction

Machine learning models form at present the core of the AI we engineer, models that tend to be complex. The field of Explainable AI (XAI) thus aims to generate explanation for such models. For reasons such as: 1) facilitate understanding in the humans collaborating with such models [37]; 2) aid the detection of biased views in a model the AI uses [131, 132]; 3) help identify situations in which the model works adequately and responsibly [133, 134, 135]; 4) offer novel insights in an underlying causal phenomena the model learned [32]; and 5) as a tool to let engineers build better models and debug existing ones [136, 82].

The existing methods in Explainable AI (XAI) focus on different approaches to obtain the information needed an explanation. See for example for an overview the review papers of Guidotti et al. [39] and Chakraborty et al. [137]. A number of examples of common methods are; ordering the features' attributions on a decision [138, 139, 140], saliency an attention feature maps [141, 142, 143, 144], prototype selection, construction and presentation [145], word annotations [86, 146], and summaries with decision trees [147, 148, 149] or decision rules [150, 151, 152, 153]. In this chapter we focus on explaining the role and importance of features on a single decision made by an AI given a tabular feature vector. Such explanations tend to be long when based on all features or use an arbitrary cutoff point to reduce the number of features in the explanation. Instead, we propose a model-agnostic method to limit the explanation length with the help of contrastive explanations. Not only offers a natural way to prevent information overload, it also aims to explain the role the relevant features had, summarized as decision rules.

Throughout this chapter, the main reason for explanations is to offer understanding in a machine learning model underlying an AI's decision making or support (i.e., a classification task). An understanding on which features played a role in a decision and what that role was. A few methods that offer similar explanations are LIME [140], QII [138], STREAK [154] and SHAP [155]. Each of these approaches answers the question "Why this decision?" in some way by quantifying the attribution to a single decision, either as a subset of features or an ordered list of all features. However, when humans answer such questions to each other they tend to limit their explanations to only few vital aspects [156]. Huysmans et al. [157] argues for something similar in XAI; when different explanations apply we should pick the simplest explanation that is consistent with the data. The aforementioned approaches do this by either thresholding the contribution parameter to a fixed value, presenting all features as an ordered list or by applying it only to low-dimensional data.

This study offers a more human-like way of limiting the contributing features by setting a contrast between two decisions. The proposed contrastive explanations present only the information that causes some data point to have one decision instead of another [59]. Recently, Dhurandhar et al. [158] proposed constructing explanations by finding contrastive perturbations—minimal changes required to change a current decision to an arbitrary other decision. Instead, our approach creates contrastive *targeted* explanations by first defining the alternative decision of interest to a human. In other words, our generated contrastive explanations answer the question "Why this decision instead of that decision?". The contrast is made between the *fact*, the current decision, and the *foil*, the decision of interest.

A relative straightforward way to construct contrastive explanations given a foil can be based on solely feature attributions. First the feature attributions for the fact and foil decisions are computer, resulting in two ordered feature lists associated to each. Then the ranking of a feature in each lists can be compared and used to explain why one decision was made instead of another; because some features play a larger role and other a smaller role. However, this does not explain what that role was and features may have the same rank in both lists but have entirely different roles. We propose a more meaningful comparison by comparing the approximate decision rule associated to each feature to decide for either the foil or fact. To do so, we train a surrogate model to distinguish between fact and foil in a more localized area of the input space. From that model we distill two sets of rules; one used to identify data points as a fact and the other to identify data points as a foil. Given these two sets, we subtract the factual rule set from the foil rule set. This relative complement is used to construct our contrastive explanation. See Figure 3.1 for an illustration.

Our proposed method obtains this complement by training a one-versus-all decision tree to recognize the foil within a localized area of the input space around the current data point. We refer to this decision tree as the Foil Tree. Next, we identify the fact-leaf—the leaf in which the current data point resides. Followed by identifying the foil-leaf, which is obtained by searching the tree with some strategy. Currently our strategy is simply to choose the closest leaf to the fact-leaf that classifies data points as the foil. The complement is then the set of decision nodes (representing rules) that are a parent of the foil-leaf but not of the fact-leaf. Rules that overlap are merged to obtain a minimum coverage rule set. The rules are then used to construct our explanation.

The method is discussed in more detail in Section 3.2. An example of its usage

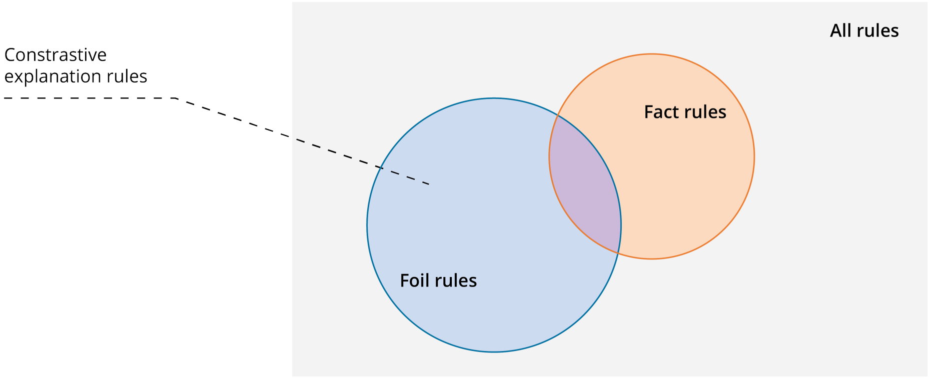


Figure 3.1: This figure illustrates our general idea of towards meaningful and targeted contrastive explanations. Given a set of rules that define data points as either the fact or foil, we take the relative complement of the fact rules in the foil rules to obtain a description how the foil differs from the fact in terms of features and decision rules.

is discussed in Section 3.3 on three benchmark classification tasks. The validation on these three tasks show, compared to other methods, that the proposed method constructs shorter explanations, provides more information on a feature's role and that this contribution matches the underlying model closely.

3.2 Foil Trees; a way for obtaining contrastive explanations

Our method learns a decision tree centred around any queried data point, commonly the current data point. The decision tree is trained to locally distinguish the foil-decision from any other decision, including the fact. We thus aim to reveal the decision boundary of the original AI through a one-versus-all classification task between foil and non foil decisions. We solve this task by training a decision tree, which allows a relatively straightforward description of the decision boundary in terms of decision rules.

The tree's training occurs on data points that can either be generated or sampled from an existing data set, each labeled with predictions from the model it aims to explain. As such, our method is model-agnostic through the use of a surrogate model. Similar to LIME [140], the sample weights of each generated or sampled data point depend on its similarity to the data point in question. With a Euclidean similarity, samples near the queried data point in the Euclidean feature space receive higher weights in training the tree. The purpose to this is to ensure the trained decision tree is more likely to be faithful to the AI's actual decision making closer to the queried data point.

Given this tree, the ‘Foil Tree’, we search for the leaf in which the queried data point resides, the so called ‘fact-leaf’. This gives us the set of rules that defines that data point not belonging to the foil. These rules respect the decision boundary of the underlying model governing the AI’s decision making, as it is trained to reflect these decisions albeit locally. Next, we use an arbitrary strategy to locate the ‘foil-leaf’—for example the leaf that classifies data points as the foil as close as possible to the fact-leaf within the tree. This results in two sets of decision rules, whose relative complement define how the data point in question differs from the foil data points as classified by the foil-leaf.

In summary, the proposed method goes through the following steps to obtain a contrastive explanation:

1. **Retrieve the fact**; the decision of the queried or current data point.

2. **Identify the foil**; the decision explicitly given in the query or otherwise derived (e.g. second most likely decision).

3. **Generate or sample a local data set**; either randomly sampled from an existing data set, generated according to a normal distribution, marginal distributions of feature values or similar.

4. **Train a decision tree**; in a weighted fashion based on a training point’s similarity to the queried data point.

5. **Locate the ‘fact-leaf’**; the leaf in which the queried data point resides.

6. **Locate a ‘foil-leaf’**; the leaf that classifies data points as part of the foil with the lowest number of decision nodes between it and the fact-leaf.

7. **Compute differences**; obtain the difference between the rules leading to the fact- and foil-leaf. Rules regarding the same feature are combined to form a single literal.

8. **Construct explanation**; the presentation of the differences between the fact- and foil-leaf.

Figure 3.2 illustrates these steps. The search for the appropriate foil-leaf in step 6 can vary. In Section 3.2.1 we discuss this more in detail. Finally, note that the method is not symmetrical. There will be a different answer on the question “Why decision A instead of B?” than on “Why decision B instead of A?” as the Foil Tree is trained in the first case to classify decision B and in the second case to classify decision A. This is because we treat the foil as the the decision of interest

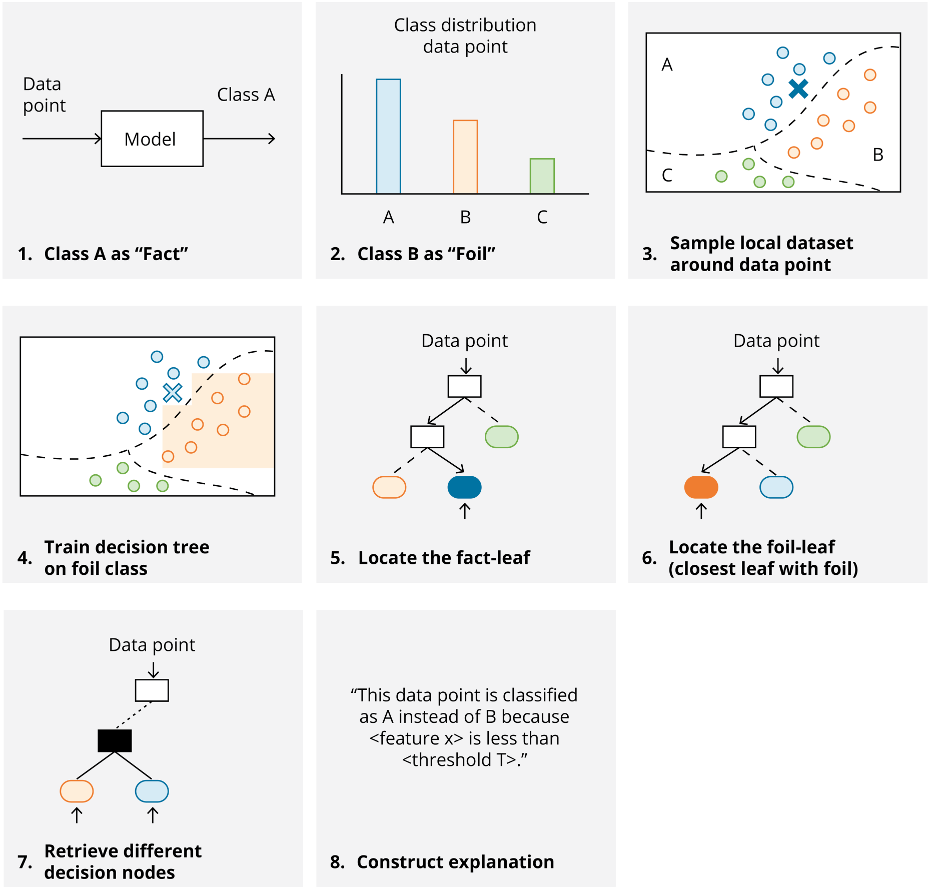

Data point
Model
Class A
1. Class A as "Fact"
Class distribution data point
A
B
C
2. Class B as "Foil"
A
B
C
3. Sample local dataset around data point
4. Train decision tree on foil class
Data point
5. Locate the fact-leaf
Data point
6. Locate the foil-leaf (closest leaf with foil)
Data point
7. Retrieve different decision nodes
"This data point is classified as A instead of B because <feature x> is less than <threshold T>."
8. Construct explanation

Figure 3.2: The steps needed to define and train a Foil Tree and to use it to construct a contrastive explanation. Each step corresponds with the listed steps in Section 3.2.

to which we compare everything else. Even if the trees end up similar, their relative complements will differ.

3.2.1 Foil-leaf strategies

We mentioned one strategy to find a foil-leaf, however multiple strategies are possible—although not all strategies may result in a satisfactory explanation according to the human. The strategy used in this study is simply the first leaf that is closest to the fact-leaf in terms of number decision nodes, resulting in a minimal length explanation.

A disadvantage of this strategy is its ignorance towards the value of the foil-leaf compared to the rest of the tree. The nearest foil-leaf may be a leaf that classifies only a relatively few data points or classifies them with a relatively high error rate. To mitigate such issues the foil-leaf selection mechanism can be generalized to an acyclic graph-search from a specific (fact) vertex to a different (foil) vertex while minimizing edge weights. The Foil Tree is treated as a graph whose decision node and leaf properties influence some weight function. This generalization allows for a number of strategies, and each may result in a different foil-leaf. The strategy used in this preliminary study simply reduces to each edge having a weight of one, resulting in the nearest foil-leaf when minimizing the total weights.

As an example, an improved strategy may be where the edge weights are based on the relative accuracy of a node or leaf. Where a higher accuracy results in a lower weight, allowing the strategy to find more distant, but more accurate, foil-leaves. This may result in relatively more complex and longer explanations, which nonetheless hold in more general cases. For example the nearest foil-leaf may only classify a few data points accurately, whereas a slightly more distant leaf classifies significantly more data points accurately. Given the fact that an explanation should be both accurate and fairly general, this proposed strategy may be more beneficial in some cases [159].

Note that the proposed method assumes the knowledge of the used foil. In all cases we take the second most likely decision as our foil. Although this may be an interesting foil it may not be the contrast a human actually wants to make. Either the human makes its foil explicit or we introduce a feedback loop in the interaction that allows our approach to learn which foil is asked for in which situations. We leave this for future work.

3.3 Validation

The proposed method is validated on three benchmark classification tasks from the UCI Machine Learning Repository [160]; the Iris data set, the PIMA Indians Diabetes data set and the Cleveland Heart Disease data set. The first data set is a well-known classification task of plants based on four flower leaf characteristics with a size of 150 data points and three classes. The second data set is a binary classification task whose task is to correctly diagnose diabetes and contains 769 data points and has nine features. The third data set is the classification of four risk categories for heart disease, consisting of 297 data points and thirteen features.

To show the model-agnostic nature of our proposed method we applied four distinct classification models to each data set; a random forest, a logistic classifier, a support vector machine (SVM) and a neural network. Table 3.1 shows for each data set and classifier the F_1 score of the trained model. We validated our approach on four measures; explanation length, accuracy, faithfulness and computation time. These measures for evaluating XAI decision rules are adapted from Craven and Shavlik [159], where the mean length serves as a measure for a rule-based explanation's interpretability [33]. The faithfulness allows us to state how well the tree explains the underlying model, and the accuracy tells us how well its explanations generalize to unseen data points. Below we describe each in detail:

1. **Length**; Mean length of the explanation in terms of decision nodes. The ideal value is in the range $[1.0,$ Nr. features$)$. A length of zero would mean that no explanation could be found. A length near or equal to the number of features could become a complex and difficult to interpret explanation.
2. **Accuracy**; F_1 score of the Foil Tree for its binary classification task on a test set compared to the true labels. Higher values indicate a method capable of generating explanations that apply to unseen data.
3. **Faithfulness**; F_1 score of the Foil Tree on the test set compared to the model's outcome. This measure provides a quantitative value of how well the Foil Tree agrees with the underlying classification model it tries to explain. Higher is better.
4. **Time**; Mean number of seconds needed to generate and explain an unseen data point. For explanations to be of practical use, they need to be generated in a feasible amount of time.

Each measure is cross-validated three times to account for randomness in a Foil Tree's construction. The results are shown in Table 3.1. They show that on average a Foil Tree is able to provide concise explanations, with a mean length of 1.33, while accurately mimicking the decision boundaries used by the model with a mean faithfulness of 93% and generalizes well to unseen data with a mean accuracy of 92%. The Foil Tree performs similar to the underlying ML model in terms of accuracy. Note that for the Diabetes data set, the random forest, logistic classification and SVM models, found explanations of length zero. In such a case, differences in rule sets could be found, which resulted in a mean length of less than one. For all other models our method was able to find a difference, and thus explanation, for every queried data point.

To further illustrate the proposed method, below we present a single explanation of two classes of the Iris data set in a dialogue setting;

Data set	Model	F_1 Score	Length	Accuracy	Faith-fulness	Time
Iris	RF	0.93	1.94 (4)	0.96	0.97	0.014
	LR	0.93	1.50 (4)	0.89	0.96	0.007
	SVM	0.93	1.37 (4)	0.89	0.92	0.010
	NN	0.97	1.32 (4)	0.87	0.87	0.005
Diabetes	RF	1.00	0.98 (9)	0.94	0.94	0.041
	LR	1.00	0.98 (9)	0.94	0.94	0.032
	SVM	1.00	0.98 (9)	0.94	0.94	0.034
	NN	1.00	1.66 (9)	0.99	0.99	0.009
Heart Disease	RF	0.94	1.32 (13)	0.88	0.90	0.106
	LR	1.00	1.21 (13)	0.99	0.99	0.006
	SVM	1.00	1.19 (13)	0.86	0.86	0.012
	NN	1.00	1.56 (13)	0.92	0.92	0.009

Table 3.1: Performance of Foil Tree explanations on the Iris, PIMA Indians Diabetes and Heart Disease classification tasks. 'Length' is the mean number of rules in each explanation with the number of features in parenthesis as its upper bound. 'Time' is the mean number of seconds to generate an explanation. The model abbreviations are; Random Forest (RF), Logistic Classification (LR), Support Vector Machine (SVM) and Neural Network (NN).

- AI: The type of Iris you describe is a Setosa.
- Human: Why a Setosa instead of a Versicolor?
- AI: Because for it to be a Versicolor the petal width should be smaller and the sepal width should be larger.
- Human: How much smaller and larger?
- AI: The petal width should be smaller than or equal to 0.8cm and the sepal width should be larger than 3.3cm.

The fact is the 'Setosa' flower type, the foil is the 'Versicolor' flower type and the total length of the explanation is two decision rules. The generation of this small dialogue is based on text templates and fixed interactions for the user.

3.4 Conclusion

Current developments in Explainable AI (XAI) created new methods to answer "Why decision A?". Many XAI methods compute the amount with which each feature can be attributed to the decision A. To prevent information overload, often only a subset of features is communicated whose contribution is above a threshold, all attributions are communicated but as an ordered list, or the method apply is applied only to cases with few features.

This chapter presented a novel method answer a contrasting query of the form “Why decision A (fact) instead of decision B (foil)?” for an arbitrary data point. This served as a natural way to reduce explanation length by only explaining the decision contrast. Furthermore, instead of only conveying the features’ attributions we also aimed to convey a features role in the decision in the form of its relevant decision rule.

Our method finds the contrastive explanation by taking the relative complement of decision rules for the fact with respect to the rules for the foil. In this chapter we implemented this idea by training a decision tree to distinguish the foil from all other decisions (a one-versus-all approach). A fact-leaf is selected as the leaf in the tree in which the queried or current data point resides. Also, a foil-leaf is selected according to some strategy that classified data points as the foil. Overlapping rules are merged, and used to construct the actual explanation.

We introduced a strategy of finding a foil-leaf that minimizes explanation length. An weighted acyclic graph-search approach was suggested as a generic way to describe various strategies. In this chapter we evaluated if the proposed method is viable on three different benchmark task. In addition, the method’s faithfulness on different underlying models to show its model-agnostic capacity. The results showed that for different classifiers our method is able to offer concise explanations that accurately describe the decision boundaries of the model it explains.

3.5 Acknowledgements

We acknowledge the project ERP Applied Artificial Intelligence from TNO for funding this research.

CHAPTER 4

CONTRASTIVE EXPLANATIONS FOR REINFORCEMENT LEARNING

This chapter is adapted from; **van der Waa, J.**, van Diggelen, J., van den Bosch, K., & Neerincx, M. (2018). Contrastive explanations for reinforcement learning in terms of expected consequences. *Proceedings of the Workshop on Explainable AI, International Joint Conference of Artificial Intelligence*. The adaptations include an altered abstract, shortened introduction, and an adjusted lay-out, including the formatting of figures and tables.

Jasper van der Waa, Jurriaan van Diggelen, Karel van den Bosch, Mark Neerincx

While machine learning is used in AI agents that makes or advises on decisions, it is also used for agents to plan their behaviour, known as reinforcement learning (RL), a form of unsupervised trial-and-error learning. As the capabilities of RL agents grow, so does their complexity, which hinders human trust and acceptance. Within this chapter, we propose a method that lets RL agents generate contrastive explanations. The method aims to explain the differences in expected and exemplar state transitions and outcomes between the agent's determined policy of behaviour and a policy of interest, derived from a human query (e.g., an explicit question or interface interaction for more information). The method can be applied to the next determined action or to an entire policy of many future actions. It also accounts for a translation between the agent's perceived states and actions and conceptions of these that are interpretable to humans. This information is to support the interpretability of the eventual contrastive explanation for which we offer a textual template. A pilot human study was conducted to explore human preferences for different properties of explanation. The results indicate that humans tend to favour explanations of an entire policy of behaviour rather than of the next single action.

4.1 Introduction

Compared to other machine learning approaches such as classification, receive reinforcement learning (RL) agents little attention in Explainable AI [39, 161, 162, 150]. Even though some of the most impressive results in AI rely on such agents, such as AlphaGo defeating the reigning human champions in Go. Due to the intrinsic different nature of RL compared to other approaches, many of the explanation generating methods that are model-agnostic do not readily apply to RL.

The scarcity of methods for RL agents to explain their actions towards humans severely hampers the practical applications of RL. It also diminishes the value of RL to Artificial Intelligence [150, 163]. Take for example a simple agent within a grid world that needs to reach a goal position while evading another agent and traps. At present, even such a simple RL agent cannot easily explain why it takes the route it has learned. For instance, because the agent is only aware of numerical rewards, its coordinates in the grid and distance to the other agent. The agent has no grounded knowledge about the other 'evil' agent that tries to prevent it from reaching its goal nor has it knowledge how certain actions will effect such grounded concepts. Present day RL agents tend to learn and make use of abstract state features and rewards. This is what drives them and their success, but do not lend themselves for an explanations. Often concepts useful for a RL agent are very difficult to understand for a human. There is a need to translate the RL agent's perceptions and reasoning into a more interpretable manner and formulate an explanation on its behaviour.

Important pioneering work has been done by Hayes and Shah [78]. They developed a method for explainable Reinforcement Learning that can generate explanations about a learned policy in a way that is understandable to humans. Their method converts state feature vectors to a list of predicates by using a set of binary classification models. This list of predicates is searched to find sub-sets that tend to co-occur with specific actions. The method provides information about which actions are performed when certain predicates hold. A method that uses the co-occurrence to generate explanations may be useful for small problems, but becomes less comprehensible in larger planning and control problems. Simply due to the combinatorial explosion of predicates and action combinations. Also, the method addresses only *what* the agent does, and not *why* it acts as it does. In other words, the method presents the human with the correlations between states and the policy but it does not provide a motivation why that policy is used in terms of rewards or state transitions.

This chapter proposes a method that allows a RL agent to answer questions about its actions and policy in terms of their consequences. Other questions unique to RL are also possible, for example those that ask about the time it takes to obtain some goal. However we believe that laymen in RL are mostly interested in the expected consequences of the agent's learned behaviour and the agent's appraisal of those consequences. This information can be used as an argument why the agent behaves in some way. This would allow humans to gain insight in what information the agent can perceive from a state and which outcomes it expects from an action or state visit. Furthermore, to limit the amount of information of all consequences, our proposed method aims to support contrastive explanations [68].

Contrastive explanations are a way of answering causal 'why'-questions. In such questions, two potential items, the fact and foil, are compared to each other in terms of their causal effects on the world. Contrastive questions come natural between humans and offer an intuitive way of gathering motivations about why one performs a certain action instead of another [68]. In our case we allow the human to formulate a question of why the learned policy π_t (the 'fact') is used instead of some other policy π_f (the 'foil) that is of interest to the human. Furthermore, our proposed method translates the set of states and actions in a set of more descriptive state classes $\mathbf{C}$ and action outcomes $\mathbf{O}$ similar to the work by Hayes and Shah [78]. This allows the human to query the agent in a more natural way as well as receive more informative explanations as both refer to the same concepts instead of plain features. The translation of state features to more high-level concepts and actions in specific states to outcomes, is also done in the proposed algorithm of Sherstov et al. [164]. The translation in this algorithm was used to facilitate transfer learning within a single action over multiple tasks and domains. In our method we used it to create a human-interpretable variant of the underlying Markov Decision Problem (MDP).

For the purpose of implementation and evaluation of our proposed method, we performed a pilot study. In this study, a number of examples of various explanation forms were presented to participants to see which they preferred. For instance, two forms differed between addressing the next action or the entire policy.

4.2 Method for consequence-based explanations

The underlying Markov Decision Problem (MDP) of a RL agent consists of the tuple $\langle S, A, R, T, \lambda \rangle$. Here, S and A are respectively the set of states described by a feature vector and actions. $R : S \times A \rightarrow \mathbb{R}$ is the reward function and $T : S \times A \rightarrow Pr(S)$ the transition function that provides a probability distribution over states. λ is the discount factor that governs how future rewards are scaled. This tuple provides the required information to derive the consequences of the learned policy π_t or the foil policy π_f specified by a human's query. We refer to these sets of consequences as γ and γ_f for the fact and foil policies respectively.

Identifying the consequences γ and γ_f relies on simulating π_t and π_f by sampling the transition function T with the current state $s_t \in S$ as starting point. In the case T is not a given, one may use a separate ML model to learn in parallel to the learning process of the agent. Through this simulation, one constructs a Markov Chain of state visits under each policy π_t and π_f and can present the difference to the human as a contrastive explanation.

Through the simulation of future states with T, information can be gathered about state consequences. In turn, from the agent itself the state or state-action values for simulated state visits can be obtained to develop an explanation in terms of rewards. However, the issue with this approach is that the state features and rewards may not be easy to interpret for a human as it would consist of possibly low-level concepts and numerical reward values. To mitigate this issue we can apply a translation of the states and actions to a set of predefined state concepts and outcomes. These concepts can be designed to be more descriptive and informative for the potential human. A way to do this translation is by training a set of binary classifiers to recognize each outcome or state concept from the state features and taken action. Their training can occur during the exploratory learning process of the agent. This translation can be combined with a policy simulation to arrive at a more human-interpretable contrastive explanation.

4.2.1 A human-interpretable MDP

The original set of states can be transformed to a more descriptive set $\mathbf{C}$ according to the function $\mathbf{k} : S \rightarrow \mathbf{C}$. This is similar to the approach of Hayes and Shah [78] where $\mathbf{k}$ consists of a number of classifiers. Also, rewards can be explained in terms of a set of previously specified action outcomes $\mathbf{O}$ according to $\mathbf{t} : \mathbf{C} \times A \rightarrow Pr(\mathbf{O})$. This provides the results of an action in some state in terms of the concepts $\mathbf{O}$. For example, the outcomes that the developer had in mind when designing the reward function R. The combined translation of states and actions in outcomes is adopted from the work of Sherstov et al. [164]. In their work the transformations are used to allow for transfer learning in RL. Here however, we use them as a translation towards a more human-interpretable repres-

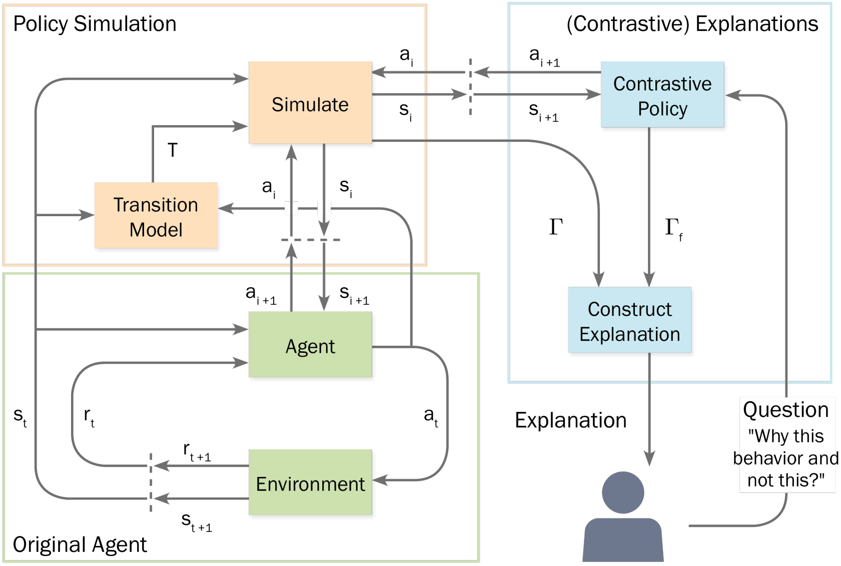


Figure 4.1: An overview of the proposed method. A dotted line represents the start of a new instance of a feedback loop. We assume a general reinforcement learning agent that acts upon a state s through some action a and receives a reward r. We continuously train a transition model T on the agent's transitions. This T is used to simulate the effect of actions on states. By repeatedly sampling a potential next state s_i from T, we can obtain the expected consequences γ of an entire policy. In a similar way we obtain the expected consequences γ_f by sampling transitions using a constrastive behaviour policy. Finally, in constructing the explanation we transform the consequences into human-interpretable concepts and construct an explanation.

entation of the actual MDP.

The result is the new MDP tuple $\langle S, A, R, T, \lambda, \mathbf{C}, \mathbf{O}, \mathbf{t}, \mathbf{k} \rangle$. An RL agent is still trained on S, A, R and T with λ independent of the descriptive sets $\mathbf{C}$ and $\mathbf{O}$ and functions $\mathbf{k}$ and $\mathbf{t}$. This makes the translations independent of the RL algorithm used to train the agent. See Figure 4.1 for an overview of this approach.

As an example take the grid world illustrated in Figure 4.2 that shows an agent in a simple myopic navigation task. The states S are the (x, y) coordinates and the presence of a forest, monster or trap in adjacent tiles with $A = \{\text{Up}, \text{Down}, \text{Left}, \text{Right}\}$. R consists of a small transient penalty, a slightly larger penalty for tiles with a forest, a large penalty shared over all terminal states (traps or adjacent tiles to a monster) and a large positive reward for the finishing state. T is skewed towards the intended result with small probabilities for the other results if possible.

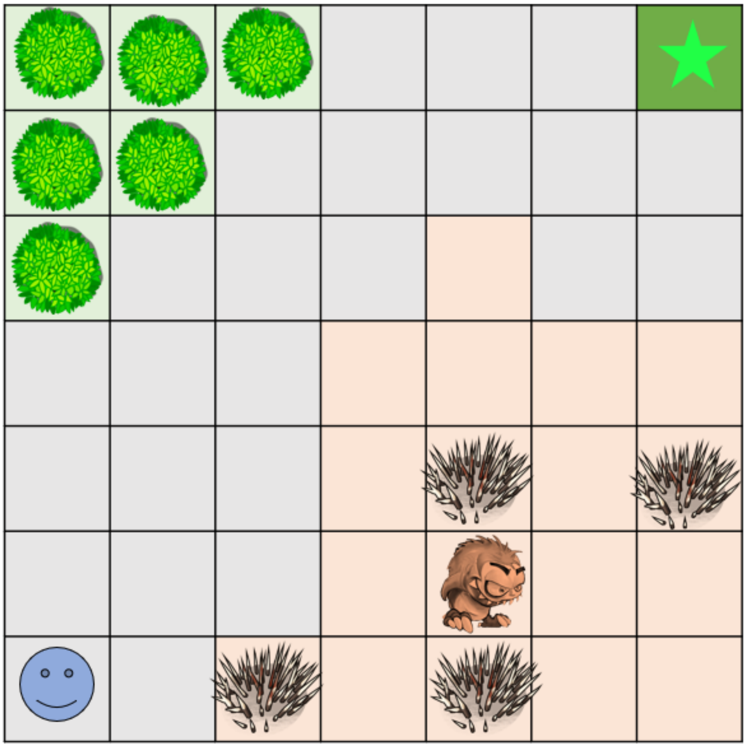


Figure 4.2: A simple RL problem where the agent has to navigate from the bottom left to the top right while evading traps, a monster and a forest. The agent terminates when in a tile with a trap or adjacent to the monster. The traps and the monster only occur in the red-shaded area and as soon as the agents enter this area the monster moves towards the agent.

The state transformation $\mathbf{k}$ can consist out of a set of classifiers for the predicates whether the agent is next to a forest, a wall, a trap or monster. For example; NearForest $(s_t = (1,3))$ or NearMonster $(s_t = (3,6))$. Applying $\mathbf{k}$ to some state $s \in S$ results in a Boolean vector $c \in C$ whose information can be used to construct an explanation in terms of the stated predicates.

The similar outcome transformation $\mathbf{t}$ may predict the probability of the outcomes $\mathbf{O}$ given a state and action. In our example, $\mathbf{O}$ could consist of whether the agent will be at the goal, in a trap, next to the monster or in the forest. Each outcome $\mathbf{o}$ can be flagged as being positive $\mathbf{o}^+$ or negative $\mathbf{o}^-$ based on the received reward. This allows them to be addressed as such in the eventual explanation.

Given the above transformations we can sample the next state of a single action a with T or even the entire chain of actions and visited states given some arbitrary policy π. These can then be transformed into state descriptions $\mathbf{C}$ and action outcomes $\mathbf{O}$ to form the basis of an explanation. As mentioned, humans usually ask for contrastive questions especially regarding their actions [68]. In the next section we propose a method of translating the foil in a contrastive question into a new policy.

4.2.2 Contrastive questions translated into value functions

A contrastive question consists of a fact and a foil, and its answer describes the contrast between the two from the fact's perspective [68]. In our case, the fact consists of the entire learned policy π_t, the next action from it $a_t = \pi_t(s_t)$ or given number of consecutive actions from π_t. We propose a method of how one can obtain a foil policy π_f derived from a human's query. An example of such a question could be (framed within the case of Figure 4.2);

> "Why do you move up and then right *(fact)* instead of moving to the right until you hit a wall and then move up *(foil)*?"

The foil policy π_f is ultimately obtained by combining a state-action value function Q_I – that represents the human's preference for some actions according to his/her question – with the learned Q_t to obtain Q_f;

$$Q_f(s,a) = Q_t(s,a) + Q_I(s,a), \forall s,a \in S \times A \tag{4.1}$$

Given that $Q : S \times A \rightarrow \mathbb{R}$.

Q_I only values the state-action pairs queried by the human. For instance, the Q_I of the above given question can be based on the following reward scheme for all potentially simulated $s \in S$;

- If 'RightWall' $\in \mathbf{k}(s)$, then $a_f^1 =$ 'Up' receives a reward such that $Q_f(s, \text{Up}) > Q_t(s, \pi_t(s))$.
- Otherwise, $a_f^2 =$ 'Right' receives a reward such that $Q_f(s, \text{Right}) > Q_t(s, \pi_t(s))$

Given this reward scheme we can obtain Q_f according to Equation (4.1). The state-action values Q_f can then be used to obtain the policy π_f using the original action selection mechanism of the agent. This results in a policy that tries to follow the queried policy as best as it can. The advantage of having π_f constructed from Q_f is that the agent is allowed to learn a different action then those in the human's question as long as it approximates it sufficiently. For instance, instead of moving to the right in trap the agent could try and evade the trap so it can move more to the right. Also, it allows for the simulation of the actual expected behaviour of the agent as it is still based on the agent's action selection mechanism. Both would not be the case if we simply forced the agent to do exactly what the human stated.

The construction of Q_I is done through simulation with the help of the transition model T. The rewards that are given during the simulation are selected with Equation (4.1) in mind, as they need to eventually compensate for the originally learned action based on Q_t. Hence, the reward for each state and queried action is as follows;

$$R_I(s_i, a_f) = \frac{\lambda_f}{\lambda} w(s_i, s_t) \left[R(s_i, a_f) - R(s_i, a_t\right] (1 + \epsilon) \tag{4.2}$$

With $a_t = \pi_t(s_t)$ the originally learned action and w being a distance based weight;

$$w(s_i, s_t) = e^{-\left(\frac{d(s_i, s_t)}{\sigma}\right)^2} \tag{4.3}$$

First, s_i with $i \in \{t, t+1, ..., t+n\}$ is the i'th state in the simulation starting with s_t. a_f is the current foil action governed by the conveyed policy by the human. The fact that a_f is taken as the only rewarding action each time, greatly reduces the time needed to construct Q_I. Next, $w(s_i, s_t)$ is obtained from a Radial Basis Function (RBF) with a Gaussian kernel and Euclidean distance function d. This RBF represents the exponentially decreasing distance between our actual state s_t and the simulated state s_i. The Gaussian kernel is governed by the standard deviation σ and allows us to reduce the effects of Q_I as we get further from our actual state s_t. The ratio of discount factors $\frac{\lambda_f}{\lambda}$ allows for the compensation between the discount factor λ of the original agent and the potentially different factor λ_f for Q_I if we wish it to be more shortsighted. Finally, $[R(s_i, a_f) - R(s_i, a_t)] (1 + \epsilon)$ is the amount of reward that a_f needs such that $Q_I(s_i, a_f) - Q(s_i, a_t) = \epsilon$. With $\epsilon > 0$ that determines how much more Q_I will prefer a_f over a_t.

The parameter n defines how many future state transitions we simulate and are used to retrieve Q_I. As a general rule $n \geq 3\sigma$ as at this point the Gaussian kernel will reduce the contribution of Q_I to near zero such that Q_f will resemble Q_t. Hence, by setting σ one can vary the number of states the foil policy should starting from s_t. Also, by setting ϵ the strength of how much each a_f should be preferred over a_t can be regulated. Finally, λ_f defines how shortsighted Q_I should be. If set to $\lambda_f = 0$, π_f will force the agent to perform a_f as long as s_i is not to distant from s_t. If set to values near one, π_f is allowed to take different actions as long as it results into more possibilities of performing a_f.

4.2.3 Generating explanations

At this point we have the human-interpretable MDP consisting of state concepts $\mathbf{C}$ and action outcomes $\mathbf{O}$ provided by their respective transformation function $\mathbf{k}$ and $\mathbf{t}$. Also, we have a definition of R_I that values the actions and/or states that are of interest by the human which can be used to obtain Q_I through simulation and Q_f according to Equation (4.1). This provides us with the basis of obtaining the information needed to construct an explanation.

As mentioned before, the explanations are based on simulating the effects with T of π_t and that of π_f. We can call T on the previous state s_{i-1} for some action $\pi(s_{i-1}$ to obtain s_i and repeat this until $i == n$. The result is a single sequence or trajectory of visited states and performed actions for any policy π starting from s_t;

$$\gamma(s_t, \pi) = \{(s_0, a_0), ..., (s_n, a_n) \mid T, \pi\} \tag{4.4}$$

If T is stochastic, multiple simulations with the same policy and starting state result in different trajectories. To obtain the most probable trajectory $\gamma^*(s_t, \pi)$ we can take the transition from T with the highest probability. Otherwise a Markov chain could be constructed through sampling instead of a single trajectory.

The next step is to transform each state and action pair in $\gamma(s_t, \pi)^*$ to the human-interpretable description with the functions $\mathbf{k}$ and $\mathbf{t}$;

$$\begin{aligned} Path(s_t, \pi) = & \{(c_0, o_0), ..., (c_n, o_n)\}, \\ & c_i = \mathbf{k}(s_i),\ o_i = \mathbf{t}(s_i, a_i), (s_i, a_i) \in \gamma^*(s_t, \pi) \end{aligned} \tag{4.5}$$

From $Path(s_t, \pi_t)$ an explanation can be constructed about the state the agent will most likely visit and the action outcomes it will obtain. For example with the use of the following template;

> "For the next n actions I will mostly perform a. During these actions, I will come across situations with $\forall c \in Path(s_t, \pi_t)$. This will cause me $\forall o^+ \in Path(s_t, \pi_t)$ but also $\forall o^- \in Path(s_t, \pi_t)$."

Let a here be the action most common in $\gamma(s_t, \pi_t)$ and both o^+ and o^- the positive and negative action outcomes respectively. Since we have access to the entire simulation of π_f, a wide variety of explanations is possible. For instance we could also focus on the less common actions;

> “For the next n actions I will perform a_1 when in situations with $\forall c \in Path(s_t, \pi_t | \pi_t = a_1)$ and a_2 when in situations with $\forall c \in Path(s_t, \pi_t | \pi_t = a_2)$. These actions prevent me from $\forall o^+ \in Path(s_t, \pi_t)$ but also $\forall o^- \in Path(s_t, \pi_t)$.”

A contrastive explanation given some question from the human that describes the foil policy π_f can be constructed in a similar manner but take the contrast. Given a foil we can focus on the differences between $Path(s_t, \pi_t)$ and $Path(s_t, \pi_f)$. This can be obtained by taking the relative complement $Path(s_t, \pi_t) \backslash Path(s_t, \pi_f)$; the set of expected unique consequences when behaving according to π_t and not π_f. A more extensive explanation can be given by taking the symmetric difference $Path(s_t, \pi_t) \triangle Path(s_t, \pi_f)$ to explain the unique differences between both policies.

4.3 Human study

The above proposed method allows a RL agent to explain and motivate its behaviour in terms of expected states and outcomes. It also enables the construction of contrastive explanations where any policy can be compared to the learned policy. This contrastive explanation is based on differences in expected outcomes between the compared policies.

We performed an online human pilot study in which 82 participants were shown a number of exemplar explanations about the case shown in Figure 4.2. These explanations addressed either the single next action or the policy. Both explanations can be generated by the above method by adjusting the Radial Basis Function weighting scheme and/or the foil’s discount factor. Also, some example explanations were contrastive with only the second best action or policy, while others provided all consequences. Contrasts were determined using the relative complement between fact and foil. Whether the learned action or policy was treated as the fact or foil, was also systematically manipulated in this chapter.

We presented the developed exemplar explanations in pairs to the participants and asked them to select the explanation that would help them most to understand the agent’s behaviour. Afterwards we asked which of the following properties they used to assess their preference: long versus short explanations; explanations with ample information versus little information; explanations addressing actions versus those that address entire behaviour policies; and explanations addressing short-term consequences of actions versus explanations that address the long-term consequences.

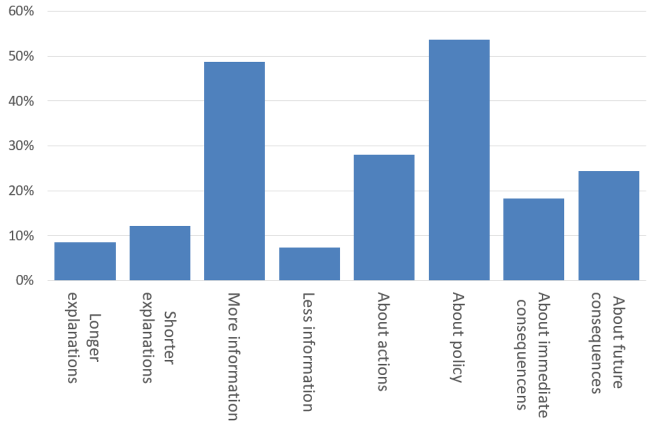


Figure 4.3: A plot depicting the percentage of participants (y-axis) preferring which explanation property (x-axis) the most over others. Answers of a total of 82 participants where gathered.

The results of the preferred factors are shown in Figure 4.3. This shows that the participants prefer explanations that address a policy that provide ample information. We note here that, given the simple case from Figure 4.2, participants may have considered an explanation addressing a single action only as trivial, because the optimal action was, in most cases, already evident to the human.

4.4 Conclusion

We proposed a method for a reinforcement learning (RL) agent to generate explanations for its actions and strategies. The explanations were based on the expected consequences of its policy. These consequences were obtained through simulation according following a state transition model. Since state features and numerical rewards do not lend themselves easily for an explanation that is informative to humans, the method supports the translation of states and actions into human-interpretable concepts and outcomes.

We also proposed a method for converting the foil, an alternative policy of interest to the human, within a contrastive 'why this instead of that?' question about actions into a policy. This simulated policy attempts to approximate the human's intent and gradually transgresses back towards its original learned policy as time progresses. How much is approximated and when the queried policy fades can each be set with a single parameter.

Through running simulations for a given number steps of both the policy derived from the human's question and the actually learned policy, we were able to obtain expected consequences of each. From here, we were able to construct contrastive explanations: explanations addressing the consequences of the learned policy and what would be different if the derived policy would have been followed.

An online survey pilot study was conducted to explore which of several explanations are most preferred by humans. Results indicate that humans prefer explanations about policies rather than about single actions.

Acknowledgments

We would like to thank the reviewers for their time and effort in improving this paper. Also, we are grateful for the funding from the RVO Man Machine Teaming research project that made this research possible.

PART II

EXPLANATIONS TO ACT UPON

CHAPTER 5

MEASURING AND EXPLAINING DECISION CONFIDENCE

This chapter is adapted from; **van der Waa, J.**, Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies, 144:102493*. The adaptations include an altered abstract, shortened introduction, and an adjusted lay-out, including the formatting of figures and tables.

Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen, Mark Neerincx

The previous chapters discussed how explanations improve human understanding and how to generate them for diverse types of AI agents. An oft-cited purpose of this understanding is that humans can determine whether the AI's decisions can be trusted. In this chapter, we propose a more specific use of explanations to support such calibration of trust: computing an AI agent's confidence in an interpretable manner and explain that computation. We define such computed confidences and their explanations as interpretable confidence measures (ICM). In two human studies we investigate what properties should define an ICM: 1) accuracy, 2) transparency, 3) explainability and 4) predictability. Case-based reasoning is presented as a method to compute such confidence measures, and exemplar implementations are proposed and evaluated on several data sets. The results show that ICM can be as accurate as common confidence measures, while behaving more predictably. The ICM's underlying idea of case-based reasoning enables an AI to explain the computation of the confidence value and facilitates a human's ability to interpret the algorithm.

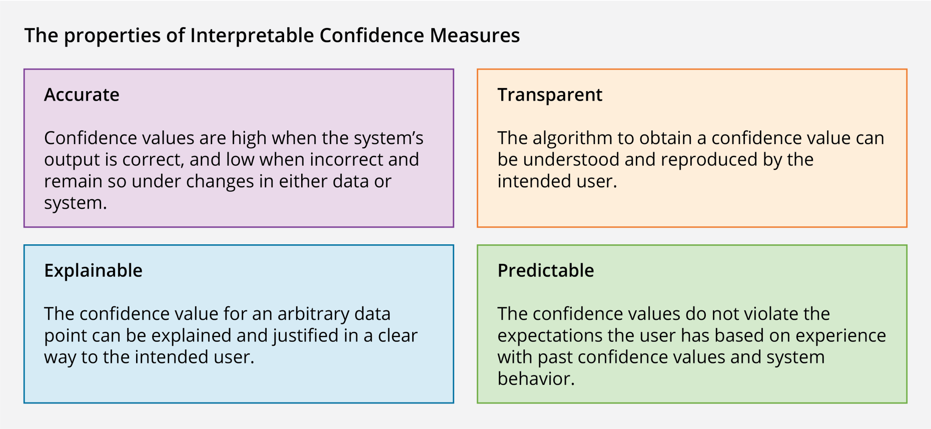
5.1 Introduction

The successes in Artificial Intelligence (AI), Machine Learning (ML) in particular, caused a boost in the accuracy and application of intelligent decision support systems (DSS). They are used in lifestyle management [165], management decisions [166], genetics [167], national security [168], and in prevention of environmental disasters in the maritime domain [169]. In these high-risk domains, a DSS could be beneficial as it can reduce human workload, and increase task performance. However, the complexity of current DSS (e.g., those based on Deep Learning) impedes a humans' understanding of a given advice, often resulting in too much or too little trust in the system, which can have catastrophic consequences [170, 171].

The field of Explainable AI (XAI) researches how a DSS can improve a human's understanding of the system by generating explanations about its behaviour [68, 59, 172, 173, 39]. More specifically, the goal of these explanations is to increase understanding of the system's rationale and certainty of an advice that it provides [126, 42, 174]. It is hypothesized that the understanding that a human gains from these explanations facilitates adequate use of the DSS [61], and calibrates the human's trust in the system [175, 176, 177].

Although understanding of the system can help humans to decide when to follow the advice of a DSS, it is often overlooked that a confidence measure can achieve the same effect [43]. In this paper, we define a *confidence measure* as a measure that provides an expectation that an advice will prove to be correct (or incorrect). To help develop such measures, we introduce the Interpretable Confidence Measure (ICM) framework. The ICM framework assumes that a confidence measure should be 1) *accurate*, 2) able to *explain* a single confidence value, 3) use a *transparent* algorithm and 4) providing confidence values that are *predictable* for humans (see Figure 5.1).

To illustrate the ICM framework, we will define an example ICM. We evaluated its accuracy, robustness and genericity on several classification tasks with different machine learning models. In addition, we applied the concept of an ICM on the use case of Dynamic Positioning (DP) within the maritime domain [169]. Here, a human operator supervises a ship's auto-pilot while receiving assistance from a DSS that provides a warning when human intervention is deemed necessary (e.g. based on weather conditions). It can be catastrophic if the operator fails to intervene in time. For example, an oil tanker might spill large amounts of oil in the ocean, because the operator failed to intervene to prevent the ship from rupturing its connection to an oil rig. This use case provided a realistic data



The properties of Interpretable Confidence Measures

Accurate

Confidence values are high when the system's output is correct, and low when incorrect and remain so under changes in either data or system.

Transparent

The algorithm to obtain a confidence value can be understood and reproduced by the intended user.

Explainable

The confidence value for an arbitrary data point can be explained and justified in a clear way to the intended user.

Predictable

The confidence values do not violate the expectations the user has based on experience with past confidence values and system behavior.

Figure 5.1: The four properties of an Interpretable Confidence Measure to perform effective trust calibration.

set to evaluate our example ICM, as well as a context for a qualitative usability study with these operators. In that study, we evaluated the transparency and explainability properties of the ICM framework. To further substantiate these results, we performed a quantitative online human study in the context of self-driving cars.

We provide the ICM framework in Section 5.3, describe our example ICM in Section 5.4, our evaluations on the data sets in Section 5.4.1, and the two human studies in Section 5.5 and 5.6. The next Section presents related work in the field of XAI and confidence measures in Machine Learning, which defines many current DSS.

5.2 Related work

Explainable AI (XAI) researches how we can improve the human's understanding in a DSS to reach an appropriate level of trust in its advice [68, 59, 37, 172, 173]. For example by allowing humans to detect biases [178, 33, 179, 180]. Some XAI research focuses on these aspects from a societal perspective, trying to identify how intelligent systems should be implemented, when they should be used, and who should regulate them [180, 32, 181, 33]. Other researchers approach the field from a methodological perspective, and aim to develop methods that solve the potential issues of applying intelligent systems in society. See for example the overview of methods from Guidotti et. al [39].

To generate explanations, many XAI methods use a meta-model that describes the actual system's behaviour in a limited input space surrounding the to be explained data point [140]. It only has to be accurate in this local space and can thus be less complex and more explainable than the actual system. A disadvantage of these approaches is that the meaningfulness of the explanation is dependent on the size of the local space and the brittleness of the used meta-model. When it is too small, the explanation cannot be generalized, and when it is too large, the explanation may lack fidelity. The advantage is that these methods can be applied to most systems (i.e. they are system- or model-agnostic). A second advantage is that the fidelity of explanations can be measured, since the meta-model's ground truth is the output of the system, which is readily available. This can be exploited to measure a meta-model's accuracy through data perturbation. In our proposed ICM framework, we apply the idea of system-agnostic local meta-models to obtain an interpretable confidence measure, not a post-hoc explanation of an output.

Confidence measures allow DSS to convey when an advice is trustworthy [43]. However, a human's commitment to follow a DSS' advice is linked to his or her own confidence and that conveyed by the DSS [182]. A confident human confronted with a low system confidence reduces the human's confidence in his or herself, and vice versa. The works from Ye et. al [88] and Waterman [183] show that this can be mitigated by explaining the DSS' confidence value by using a transparent algorithm. The work from Walley [184] shows humans tend to change their confidence when evidence for a correct or incorrect decision is gained or lost. Humans expect the same predictable behaviour from a DSS' confidence measure. Hence, it should not only be transparent with explainable values but also behave predictable for humans.

Current DSS are often based on Machine Learning (ML). Different categories of confidence measures can be identified from this field, see Table 5.1 for an overview. The first, *confusion metrics* such as accuracy and the F1-score, are based on the confusion matrix. These tend to be transparent and predictable but lack accuracy and explainability for conveying the confidence of a single advice [185, 190]. A ML model's *prediction score* such as the SoftMax output of a Neural Network, are also common as confidence measures. They represent the model's estimated likelihood for a certain prediction [191]. They are highly accurate but their transparency and explainability is often low [34, 192]. Furthermore, these measures tend to behave unpredictable as small changes in a data point can cause non-monotonic increases or decreases in the confidence value [193, 194]. In *rescaling* such as with Platt Scaling [187] or Isotonic Regression [195, 196], the prediction scores are translated into more predictable and accurate

			Category		
Property	Confusion metrics	Prediction scores	Rescaling	Probability	Voting
Accurate	-	+	+	+	+
Predictable	+	-	+	-	-
Transparent	+	-	-	-	-
Explainable	-	-	-	+	+
Example	F1-score [185]	SoftMax [186]	Platt Scaling [187]	RVM [188]	Random Forest [189]

Table 5.1: Categories and examples of commonly used confidence measures in Machine Learning and if their adherence to the four properties of an ICM.

values [197, 198]. However, these are used to enable post-processing and not intended to be explainable or transparent [199]. Some ML models are inherently *probabilistic* and output conditional probability distributions over its predictions. Examples are Naive Bayes [200], the Relevance Vector Machine [188] and using neuron dropout [201] or Bayesian inference [202, 203, 204] on trained Neural Networks. Although they are accurate, they are also opaque and difficult to predict as conditional probabilities are difficult to comprehend by humans [205, 206]. There are efforts to make such values more explainable for specific model types, see for example Qin et. al [207] and Ridgeway et. al [173]. Finally, ML models are known to use voting to arrive at a confidence value [208, 209, 210]. Known examples are Random Forest, Decision Trees and ensembles of Decision Stumps [211]. These confidence values can be explained through examples [212]. However, their algorithmic transparency depends on the model and their values tend to change step-wise given continuous changes to the input, making them hard to predict by humans.

As can be seen in Table 5.1, neither category is accurate, predictable, explainable and transparent in a DSS context. A likely reason is that the purpose of these measures is to convey performance of a ML model to a developer, not the confidence of a DSS in an advice to a human. As a consequence, many of these measures are tailored to work for a specific or subset of model types. Only the confusion metrics of these categories are system-agnostic. In the next section we propose a system agnostic approach to confidence measures based on case-based reasoning that are not only as accurate as the above described measures, but also transparent, explainable and predictable.

5.3 A framework for Interpretable Confidence Measures

In this section we propose a framework to create Interpretable Confidence Measures (ICM) that are not only accurate in their confidence assessment, but whose values are predictable as well as explainable based on a transparent algorithm. The ICM framework relies on a system-agnostic approach and performs a regression analysis with the correctness of an advice as the regressor. It does so based on case-based reasoning.

Case-based reasoning or learning provides a prediction by extrapolating labels of past cases to the current queried case [213]. The basis of many case-based reasoning methods is the k-Nearest Neighbours (kNN) algorithm [214]. This method follows a purely lazy approach [215]. When queried with a novel case, it selects the k most similar cases from a stored data set and assigns the case with a weighted aggregation of the neighbour's labels. The advantage of case-based learning methods is that its principle idea is closely related to that of human decision making [216, 217, 218]. This makes such algorithms easier to understand and interpret [219]. In addition, they allow for example-based explanations of a single prediction [220]. These properties are exploited in the ICM framework to define a confidence measure as performing a regression analysis with case-based reasoning.

5.3.1 The ICM framework

In this section we formally describe the ICM framework. We assume the DSS as a function $f : \mathbb{R}^l \to \mathbb{Y}$ that assigns an advice $y \in \mathbb{Y}$ to data points $\vec{x}$ of l dimensions. It does this with a certain accuracy relative to the ground truth or label $y^* \in \mathbb{Y}$. An ICM goes through four steps to define the confidence value $C(\vec{x})$ for $\vec{x}$: 1) an *update* step, 2) a *selection* step, 3) a *separation* step, and 4) a *computation* step. Below we discuss these steps, and an overview is shown in Figure 5.2.

In the first step, the *update*, a memory $D = \{(\vec{x}_1, y_1^*), ..., (\vec{x}_n, y_n^*)\}$ is updated. This D forms the set of cases from which the confidence is computed. Given an update procedure u and new data-label pairs $(\vec{x}, y^*)$, an ICM continuously updates this memory $D' = u\left((\vec{x}, y^*), D\right)$ such that $|D| = n$. This ensures that D adapts to changes in the DSS over time. The initial D is initialized with a training set but is expanded and replaced with novel pairs during DSS usage. The size of D is fixed to n, and maintained by u. Examples of u can be as simple as a queue (newest in, oldest out) or based on more complex sampling methods (e.g. those that take the label and data distributions into account).

In the *selection step* a set S is sampled from D such that $S = s(\vec{x}, y|D)$, where s

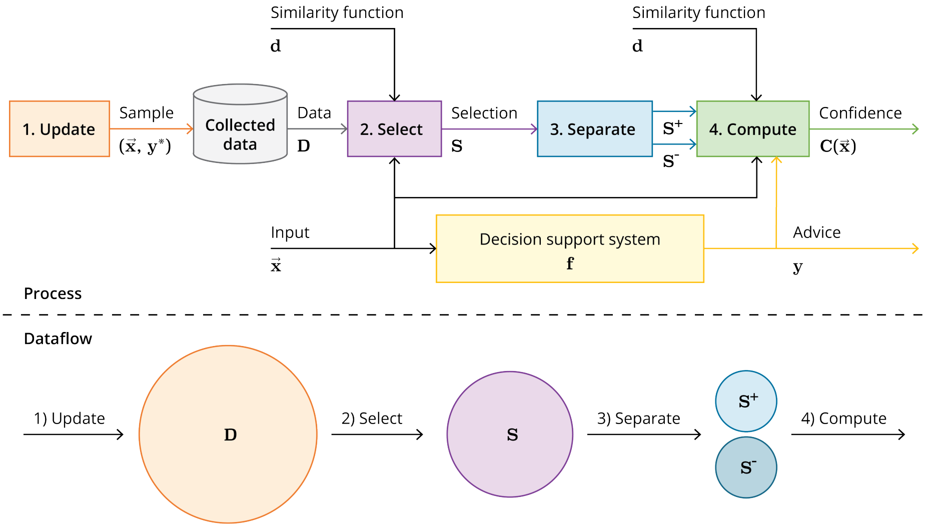


Figure 5.2: A visual depiction of the ICM framework and its four steps for computing the confidence value C for a data point $\vec{x}$ and advice y. Given a continuously updated data set D, set S is selected containing relevant data-label pairs. This S is separated in S^+ and S^- with the points that show y is correct or incorrect respectively. The confidence value $C(\vec{x})$ is computed using S^+ and S^-.

is some selection procedure. The purpose of s is to select all relevant data-label pairs to define the ICM's confidence value for the current $(\vec{x}, y)$. For example, following kNN, the k closest neighbours to $\vec{x}$ can be selected based on a similarity or distance function.

In the *separation step*, S is split into S^+ and S^- based on the current $(\vec{x}, y)$. The S^+ contains all $(\vec{x}, y^*)$ where $y = y^*$, with $S^- = S \backslash S^+$. In other words, S^+ contains all data points whose advice was similar to the current advice and correct. The S^- contains all data points with a different correct advice.

In the *computation step*, the S^+ and S^- are used to calculate the confidence value $C(\vec{x}|S^+, S^-)$ with a weighting scheme $w : \mathbb{R}^l \to \mathbb{R}$ (often abbreviated as $C(\vec{x})$:

$$C(\vec{x} \mid S^+, S^-) = Z(\vec{x}|S)^{-1} \sum_{\vec{x}_i \in S^+} w(\vec{x}, \vec{x}_i) - \sum_{\vec{x}_j \in S^-} w(\vec{x}, \vec{x}_j) \tag{5.1}$$

The weights w represent how much a data point in S^+ or S^- influences the confidence of the advice for $\vec{x}$. Again, taking kNN as an example, the w can simply contain a delta-function to 'count' the number of points in S^+ and S^-. Although, more complex weighting schemes are possible and advised. The Z^{-1} is a normalization factor:

$$Z(x)^{-1} = \frac{1}{\sum_{\vec{x}_i \in S} w(\vec{x}, \vec{x}_i)} \tag{5.2}$$

This ensures that the confidence value is bounded; $C \in [-1, 1]$, with -1 and $+1$ denoting the confidence that some y would prove to be incorrect or correct respectively. Intermediate values represent the surplus of available evidence for a correct or incorrect advice relative to all available evidence. For example, when $C(\vec{x}) = -0.5$ there is 50% surplus evidence that the advice y will be incorrect, relative to all available evidence. What constitutes as 'evidence' is determined by s to select relevant past data-label pairs and the weighting scheme w to assign their relevance. An ICM allows w and s to be any weighting scheme or selection procedure. Following other case-based reasoning methods, w and s often use a similarity or distance measure (e.g. Euclidean distance).

5.3.2 The four properties of ICM

In this section we explain why the above proposed ICM framework results in confidence measures that are not only accurate, but also predictable, transparent and explainable.

Accurate. We define the accuracy of a confidence measure as its ability to convey a high confidence for either a correct or incorrect advice, when the advice is indeed correct or incorrect. For an ICM, this can be defined as:

$$a = \frac{1}{|D|} \sum_{i=0}^{|D|} \delta\left(\vec{x}_i, y_i^*, C(\vec{x}_i)\right) \tag{5.3}$$

Where δ is the Kronecker delta, with $\delta = 1$ when $f(\vec{x}) = y^*$ and $C(x) \geq 0$, or when $f(\vec{x}) \neq y^*$ and $C(x) < 0$. Overall, case-based reasoning methods are often accurate enough for realistic data sets [221]. However, the accuracy depends on the choice for the selection procedure s and weighting scheme w. If one chooses a simple kNN paradigm, one may expect a lower accuracy then when using a more sophisticated s and w. More complex options could include learning a complex similarity measure [186]. This potentially increases the accuracy, but at the cost of ICM's transparency and predictability.

Predictable. A confidence measure should behave predictable; it should monotonically increase or decrease when more evidence or data becomes available for an advice being correct or incorrect respectively. For an ICM to be predictable, it must use a monotonic similarity function. Any step-wise or non-monotonic similarity function creates confidence values that suffer from changes that are unexpected for humans. In addition, with the update procedure u an ICM adjusts its confidence according to any changes in the data distribution or DSS itself.

Transparent. An ICM is transparent; its algorithm can be understood relatively easily by humans. Case-based reasoning is often applied by humans themselves [216]. This makes the idea of an ICM, recall past data-label pairs and extrapolate those to the current data point into a confidence value, relatively easy to comprehend. A deeper understanding of the algorithm may be possible, but depends on the complexity of the similarity measure, the selection procedure s and weighting scheme w.

Explainable. The confidence of an ICM can be easily explained using examples as selected from S^+ and S^-. It allows for a template-based explanation paradigm, for example:

> “I am $C(\vec{x})$ confident that y will be correct based on $|S|$ past cases deemed similar to $\vec{x}$. Of these cases, in $|S^+|$ cases the advice y was correct. In $|S^-|$ cases the advice y would be incorrect.”

These cases can then be further visualized through an interface, for example with a parallel-coordinates plot [222]. Such plots provide a means to visualize high-dimensional data and convey the ICM’s weighting scheme. They allow humans to identify if the selected past data points and their weights make sense and evaluate if what the ICM constitutes as evidence should indeed be treated as such. It may even enable a human to interact with the ICM by tweaking its potential hyper parameters (e.g. parameters for the selection procedure and weighting scheme).

Existing research such as that by Mandelbaum et. al [223], Subramanya et. al [224] and Papernot et. al [186] can be framed as an ICM. All are based on case-based learning and can be described by the four steps of the framework. However, their transparency and predictability tends to be limited due to their choice to use a Neural Network to define their similarity measure. This hinders the ICM’s transparency and predictability, but still allows the generation of explanations.

5.4 ICM examples

In this section we propose three examples of implementing an ICM using relatively simple techniques from the field of case-based reasoning. To define our ICM, we need to define the update procedure u, the selection procedure s and the weighting scheme w. The u remains unchanged: A queue mechanism that stores the latest $(\vec{x}, y^*)$ pair and removes the oldest from D.

The first example, ICM-1, is based on kNN and use it to define both s and w. The selection procedure is $S = s(\vec{x}|D, k, d)$ which selects the k closest neighbours in D to $\vec{x}$ with d being a distance function. The weighting scheme becomes $w(\vec{x}, \vec{x}_i) = 1, \forall \vec{x}_i \in S$. When applied to Equation (5.1), the resulting ICM counts and the relative number of points in S^+ and S^- to arrive at a confidence value:

$$C(x \mid S^+, S^-) = \frac{1}{k}\left(|S^+| - |S^-|\right) \tag{5.4}$$

This reflects the idea that confidence is ≥ 0 when the majority of k nearest neighbours are in favor of the given advice, and < 0 otherwise.

For our second example, ICM-2, we extend ICM-1 with the idea of Weighted kNN [225, 226]. It weights each neighbour with a kernel based on its similarity to $\vec{x}$ according to a distance function d. Given a Radial Basis Function (RBF) as kernel, the weighting scheme becomes $w(\vec{x}, \vec{x}_i) = \exp\left[-\left(\frac{1}{\sigma}d(\vec{x}, \vec{x}_i)\right)^2\right]$. The σ is the standard deviation of the RBF and we set it to $\sigma = d(\vec{x}, \vec{x}_{k+1})$ where $\vec{x}_{k+1}$ is the $k+1$ distant neighbour of $\vec{x}$. If we choose to use the Euclidean distance $d = ||\vec{x} - \vec{x}_i||^2$, the confidence value becomes:

$$\begin{aligned} C(x|S^+, S^-, d) = & \frac{1}{|S^+|}\sum_{\vec{x}_i \in S^+} \exp\left[-\left(\frac{1}{\sigma}||\vec{x} - \vec{x}_i)||^2\right)^2\right] - \\ & \frac{1}{|S^-|}\sum_{\vec{x}_j \in S^-} \exp\left[-\left(\frac{1}{\sigma}||\vec{x} - \vec{x}_j)||^2\right)^2\right] \end{aligned} \tag{5.5}$$

These values depend not only on the number of points in S^+ and S^-, but also on their similarity to $\vec{x}$. With this RBF kernel neighbours are weighted exponentially less important as they become dissimilar to $\vec{x}$.

In our third example, ICM-3, we build further on ICM-2. In it, we estimate σ for each confidence value as the average similarity between the k neighbours. Hence, ICM-3 remains equal to Equation (5.5), instead with $\sigma = \frac{1}{k}\sum_{\vec{x}_i \in S} ||\vec{x} - \vec{x}_i||^2$. With it, ICM-3 provides confidence values that take the number of data points in S^+ and S^- into account, but also weighs their similarity to $\vec{x}$ according to how similar the k neighbours are to each other. Meaning that the neighbour most similar to $\vec{x}$ contributes the most to the confidence estimation relative to the other $k-1$ neighbours.

5.4.1 Comparison of exemplar ICM behaviour

In this section we evaluate ICM-1, ICM-2 and ICM-3 and assess their behaviour, accuracy and predictability over changes in the data.

See Table 5.2 for the confidence values of all three example ICM on a synthetic 2D binary classification solved by standard SVM. This data set was generated using Python's SciKit Learn package [227]. The table contains six plots of ICM-1, ICM-2 and ICM-3 with $k = 2$ and $k = 8$. ICM-1 shows a high confidence when we would expect it. As points with a certain prediction approaches memorized points (in Euclidean space) with that prediction as their label, the confidence for a correct predictions increases. As opposed to an increasing confidence for an incorrect prediction when such points approach memorized data points whose label is different than the prediction. ICM-1 does show abrupt confidence changes with $k = 2$, that decrease for $k = 8$. Similar behaviour can be seen for ICM-2 and ICM-3. The difference is that both show even smaller abrupt changes due to their RBF kernel, with ICM-3 being the smoothest as the kernel adapts to the local density. For $k = 8$ we see that ICM-2 and ICM-3 result in an overall lack of confidence. With higher k values, S starts to contain nearly all data points from D. The summed weights for S^+ and S^- begin to represent the label ratio and confidence goes to zero. This sensitivity is likely unique to our ICM examples, and state of the art case-based reasoning algorithms are less likely to be as sensitive to k or use a different mechanism than kNN.

Next, we evaluate the accuracy of ICM-3 on two benchmark classification tasks each solved by a Support Vector Machine (SVM), Random Forest and Multi-layer Perceptron (MLP). We chose for ICM-3 as the most sophisticated ICM example. The confidence accuracy of ICM-3 was computed using Equation (5.3). The confidence values of the SVM were computed using Platt scaling [187], of the Random Forest using its voting mechanism, and of the MLP by setting SoftMax as its output layer's activation function. Since neither of these confidence values could express a high confidence for an incorrect classification, the accuracy from Equation (5.3) was adjusted to measure zero confidence as correct for an incorrect classification. The two classification tasks were a handwritten digits recognition task [228] and the diagnoses of heart failure in patients [229]. The data set properties, trained models and their hyper parameters are summarized in Tables Table 5.3 and Table 5.4 respectively.

Figure 5.3 shows the results from ten different runs per test set and model combination. The ICM performs equally well in confidence estimation as the models on both data sets. It shows that an ICM can be applied to a variety of models and performs equally well in terms of estimating when a classification

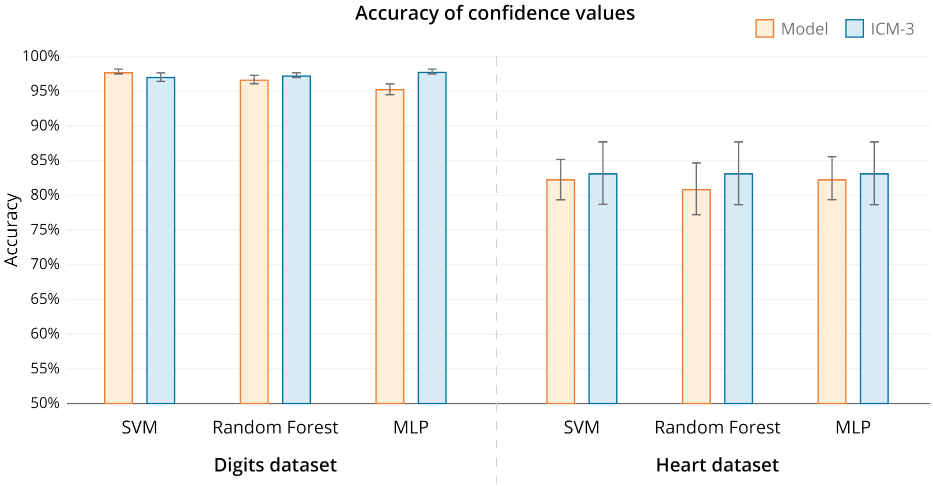


Figure 5.3: The accuracy of ICM-3 on two data sets with the accuracy of the confidence estimates from various models. It shows that ICM implementations are applicable to different models and can be equally accurate as the model itself. The error bars represent the 95% confidence intervals.

would be correct. In addition, an ICM conveys also its confidence in a classification being incorrect and tends to be more transparent, predictable and explainable.

Figure 5.4 shows the accuracy of the example ICM over different values for k. The n was set to encapsulate the entire training data set. This figure shows that ICM-3 is most robust against different values for k. More state of the art algorithms based on kNN can be applied to increase this robustness further, or algorithms based on an entirely different paradigm can be used to define the selection procedure s.

Figure 5.5 shows how the example ICM behaves with different numbers of memorized data points. The k was fixed to its optimal value of 10 neighbours for both data sets. These results show that even these simple ICM are accurate at memory sizes around 10% of the data the models needed for training.

In a separate study [169] we applied ICM-3 to a real-world DSS in the Dynamic Positioning case described in the introduction. This case was also used in one of our human studies, described in detail in Section 5.5. Here, a Deep Neural Network predicted when an ocean ship was likely to drift of course and notified a human operator to intervene. The ICM-3 was used to express more information to the operator on whether a prediction could be trusted to prevent under- or over-trust. In this study, we showed that ICM was able to compute a confidence

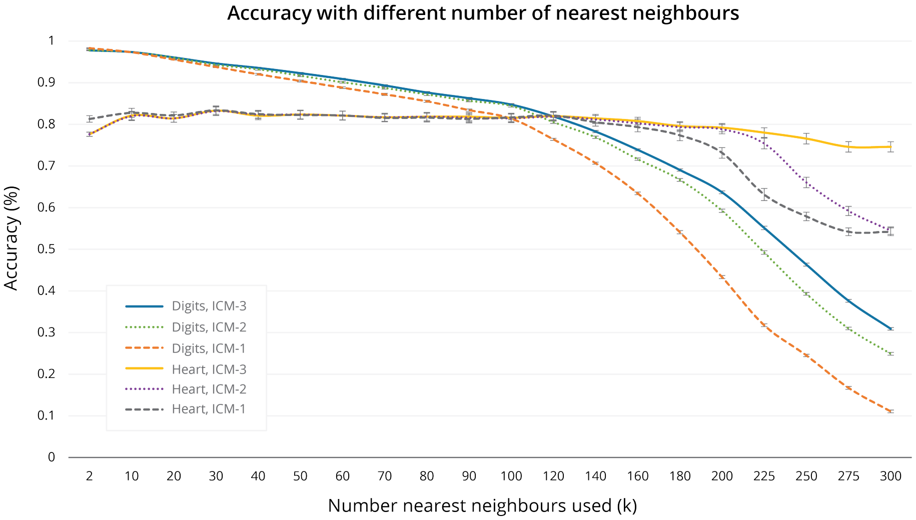


Figure 5.4: The accuracy of ICM on two data sets for different numbers of nearest neighbours used. It shows our proposed Robust Weighted kNN algorithm (ICM-3) compared to ICMs with weighted kNN (ICM-2) and kNN (ICM-1). It illustrates the robustness of ICM-3 against different k.

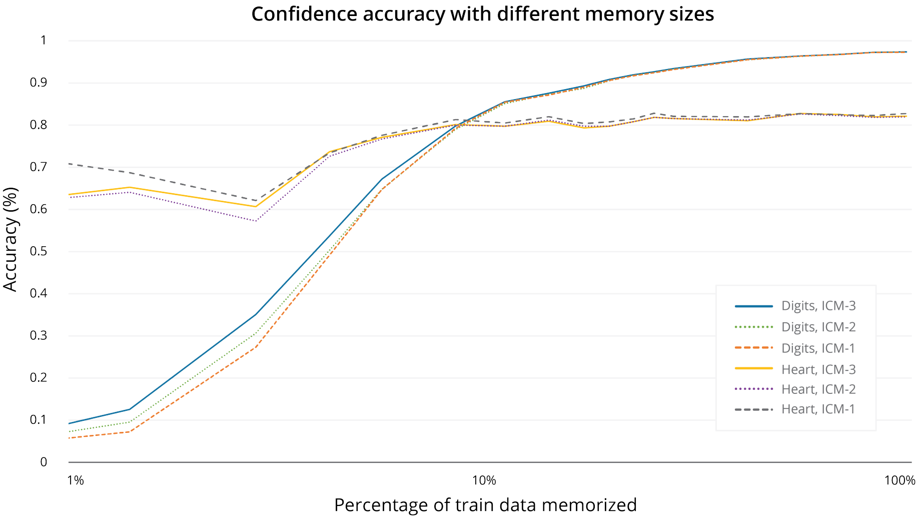


Figure 5.5: The accuracy of the example ICM on the two benchmark data sets for different values of n, the number of memorized data points. It illustrates the robustness of each ICM with different n.

value of the Deep Neural Networks prediction with 87% accuracy [169].

Finally, we evaluated how well ICM-3 could to adjust its confidence values to a

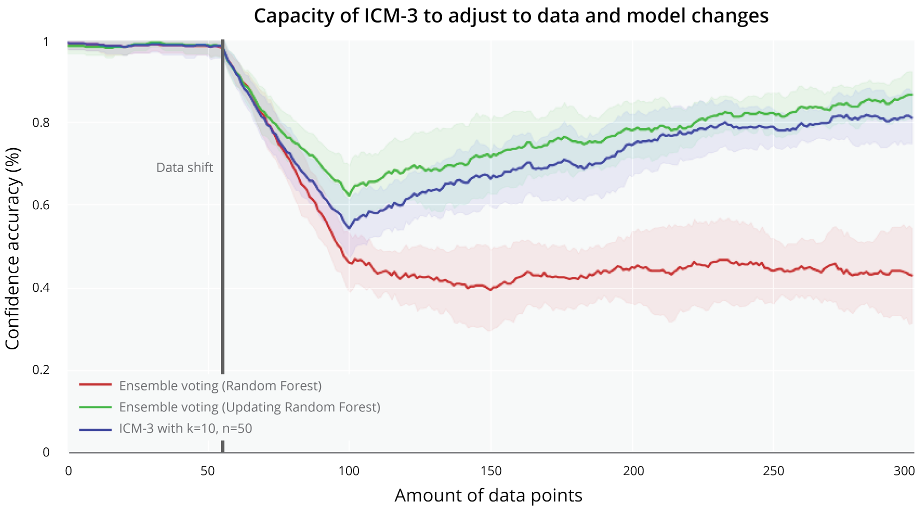


Figure 5.6: The moving average accuracy of querying confidence values from ICM-3, a static Random Forest and a continuously updated Random Forest. It shows a shift in the label distribution after 100 data points in the synthetic data. The plot shows that ICM-3 is capable of adjusting its confidence values nearly as well as the confidence from the continuously updating model.

shift in data and label distributions. As stated in Section 5.4, the update procedure u used a simple queuing method to update D. To test the effects this u has on the accuracy, we synthesized a non-linear classification task and shifted its distributions after computing the confidence of 100 data points. We compared this confidence accuracy over time with the performance of the Random Forest model with and without continuously updating that model.

The results of are shown in Figure 5.6, repeated ten times with different random seeds to obtain the shown confidence bounds. The plot shows that ICM-3 can adjust its confidence estimation to abrupt changes in the data distribution. It performs nearly equal to continuously retraining the model for each new data point, however the ICM requires no explicit update.

The results illustrate that an even simple ICM can perform surprisingly accurate on two benchmark data sets and different models. Even a simple update of the memorized data points result in an confidence estimation adaptive to changes in the data. It shows that ICM can provide a common framework to devise system-agnostic confidence measures.

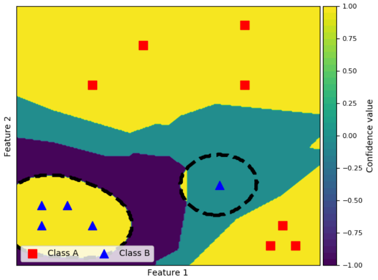
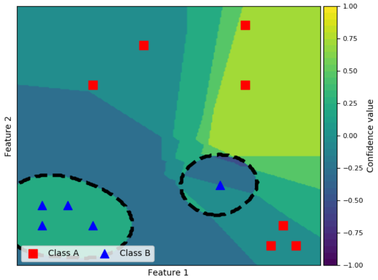
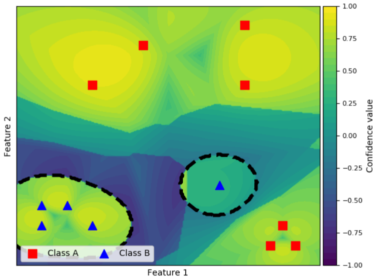
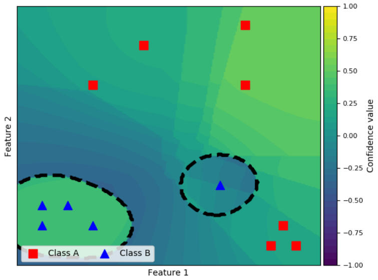
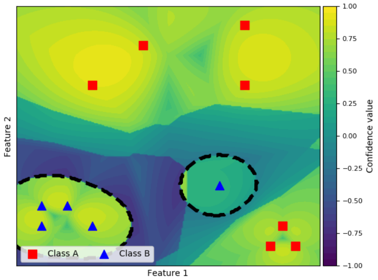
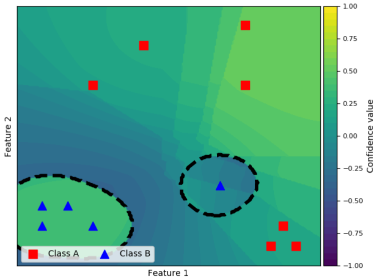
	$k = 2$	$k = 8$
ICM-1		
ICM-2		
ICM-3		

Table 5.2: These figures show the confidence values for the three example ICM implementations on a 2D synthetic binary classification task for $k = 2$ and $k = 3$. The background of each figure represents the confidence value at that point, the classification model's decision boundaries are shown by the dashed lines and D is plotted as points coloured by their true class label.

Name	Classes	Type	Features	Train/test
Heart [229]	3	Tabular	4	227/76
Digits [228]	10	Images	64	1347/450
Synthetic	6	Tabular	2	100/300

Table 5.3: The properties of the two benchmark data sets used to evaluate the three ICM examples. Also shows the properties of the synthetic data used to evaluate the robustness to changes in data distributions.

Data	Model	Accuracy (train)	Parameters
Heart	SVM	77.63% (94.27%)	RBF kernel, $\gamma = 0.1$, $C = 1.0$
Heart	MLP	76.32% (91.85%)	Adam optimizer ($\alpha = 1e^{-3}$, decay$= 1e^{6}$), 100 epochs, 16 batch size, Softmax output layer, 2 hidden Layers ([16, 3], ReLu, 20% dropout)
Heart	Random Forest	73.68% (100%)	Gini, Bootstrapping, 50 estimators
Digits	SVM	98.22% (100%)	RBF kernel, $\gamma = 0.01$, $C = 10.0$
Digits	MLP	99.33% (99.73%)	Adam optimizer ($\alpha = 1e^{-4}$, decay$= 1e^{6}$), 250 epochs, 16 batch size, Softmax output layer, 3 hidden layers ([64, 32, 6], ReLu, 20% dropout)
Digits	Random Forest	97.33% (100%)	Gini, Bootstrapping, 100 estimators
Synthetic	Random Forest	97.33% (100%)	Gini, Bootstrapping, 100 estimators

Table 5.4: Shows the hyper parameters and accuracy on train- and test set for each model and data set combination used to compare our example ICM with. We used the SciKit Learn package from Python as the implementation of each model [227].

5.5 A qualitative human study: Interviews with domain experts

This section summarizes the first of two human studies, it is explained in more detail in our previous work [230]. In this study, several domain experts were interviewed to evaluate the transparency of the case-based reasoning approach underlying an ICM compared to other confidence measures.

Dynamic Positioning (DP) formed the use case of the study. Here, a ship's bridge operator is responsible for maintaining the ship's position aided by an auto-pilot and a DSS [169]. The DSS warns the operator when the ship's position is expected to deviate from course and operator intervention is required. Structured interviews with DP operators were conducted where we elicited their understanding and needs of a confidence value that accompany the DSS' prediction. Three confidence measure categories were evaluated; 1) ICM, 2) Platt Scaling and 3) SoftMax activation functions.

The interview was structured in three phases. In the first phase we provided a layman's - but complete - description of each confidence measure. Participants were asked to select their preferred method followed by explaining each measure in their own words. This enables us to discover which algorithm they preferred, but also which they could reproduce accurately (signifying a better understanding). We found that they understood ICM best, but preferred the SoftMax measure. When asked, participants mentioned that estimating confidence in their line of work is difficult and as such they expected a confidence measure to be very complex. This result points towards what humans might prefer in a confidence measure (complexity), may not necessarily be what they need (transparency).

The second phase provided examples of realistic situations, the DSS' prediction and a confidence value. Each example was accompanied by three explanations, one from each measure. Participants were asked which explanation they preferred for each example. On average, they preferred the explanations from ICM as it specifically addressed past examples and explained their contribution to the confidence value. Afterwards, participants were asked to explain in their own words how each confidence measure would compute their values for unseen situations. The results showed that the operators could replicate ICM's explanations more easily than that of the other two.

The third and final phase allowed the participants to describe their ideal confidence measure for the DSS. Several participants described a case-based reasoning approach as used by ICM. Others preferred a combination of both an ICM and SoftMax. When asked why, they replied that they preferred the case-based reasoning approach but they believed it to be too simplistic on its own to be accurate in their line of work. They tried to add their interpretation of a SoftMax activation function to ICM to satisfy their need for added complexity.

These results may indicate that domain experts are able to understand a case-based reasoning approach for a confidence measure more easily than the DSS' prediction scores defined by a SoftMax output layer, or the scaled prediction scores with Platt Scaling.

5.6 A quantitative human study: An online survey on human preferences

The second study was performed using a quantitative online survey. We evaluated humans' interests and preferences concerning explanations about the confidence of an advice as provided by a DSS. Moreover, we investigated if the proposed ICM, based on case-based reasoning, is in line with what humans desire from a confidence measure and explanations.

Below we describe the use case, participant group, stimuli, design and analyses in more detail, followed by the results.

5.6.1 Use case: Autonomous driving

In the survey, participants were provided a written scenario about an autonomous car. This scenario stated that the car could provide an advice to turn its self-driving mode on or off, given the current and predicted road, weather and traffic conditions. The advice would be accompanied by a confidence value as calculated by the car. Participants were instructed to assume several years of experience with the car and that the car showed to be capable of driving autonomously on frequently used roads. At some point on such a familiar road, the car would provide the advice to turn on automatic driving mode. The study followed with a questionnaire revolving around this advice and the given confidence value.

5.6.2 Participants

Recruitment was done via Amazon's Mechanical Turk, and each participant received \$0.45 for participating in the survey based on the estimated time for completion and average wages. Only participants of 21 years or older were included. A total of 26 men and 14 women aged between 24 and 64 years (M=35.6, SD=9.4) were recruited, who were all (self-rated) fluent English speakers. On average, participants indicated on a 5-point Likert scale that they had some prior knowledge with self-driving cars (M=3.00, SD=0.68). Hence, participants could be biased towards answering questions based upon knowledge about self-driving cars, instead of using the description in the questions. However, the scores on the dependent variables of the participants indicated they were knowledgeable (n=6) or very knowledgeable (n=1) did not significantly differ from the scores of others and were included.

5.6.3 Stimuli

We composed a survey in which participants were asked about their interests and preferences concerning explanations about the confidence of an advice as provided by a self-driving car. The system was presented as being able to drive perfectly without assistance from a human within most situations, but unable to drive fully autonomously in some other undefined situations. We asked participants to indicate how much importance they would attach to: 1) understanding the confidence measure's underlying algorithm, 2) their past experience with other advice from the car, and 3) predictions about future conditions such as the weather. The importance was indicated on a 7-point Likert scale with 1 meaning 'not at all important' and 7 meaning 'very much important'.

Moreover, we asked participants to rank five methods of presenting the advice that the car could provide (with 1 being most preferred, and 5 being least preferred):

a. No additional information;

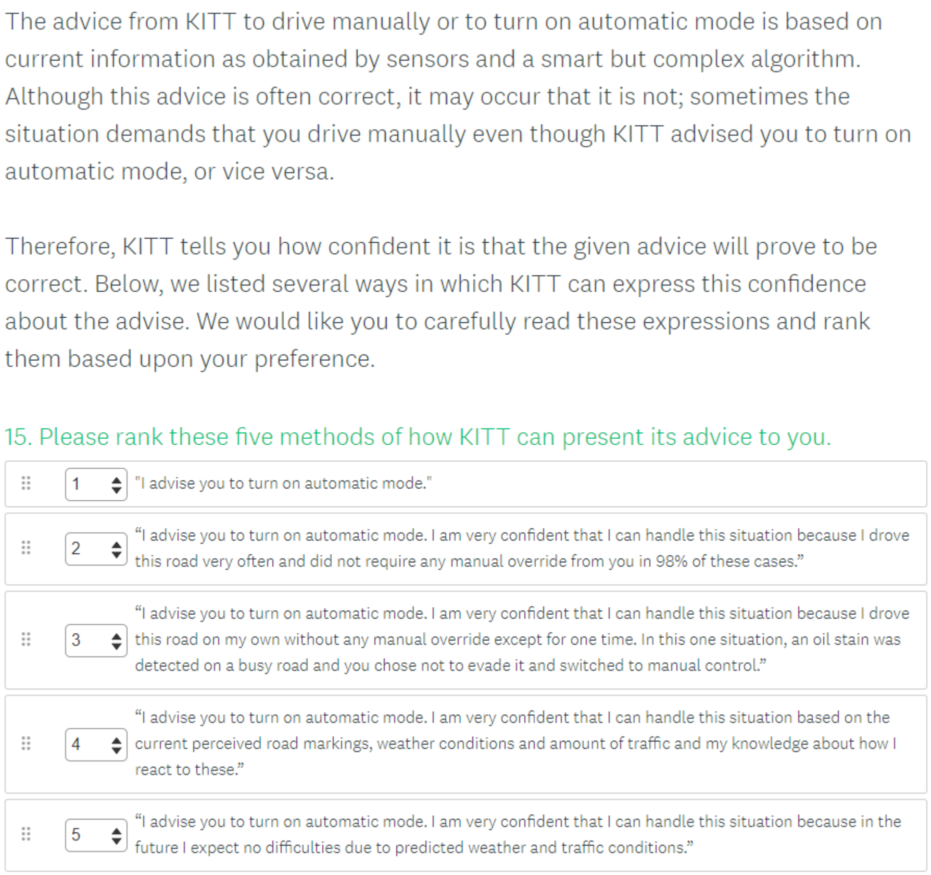
b. A general summary of prior experiences;

c. General prior experience accompanied by an illustrative specific past experience;

d. Current situational aspects that played a role;

e. Predicted future situational characteristics that could affect the decision's outcome.

Figure 5.7 shows a screenshot that contains the question in that asked participants to rank different types of explanations according to their preference. Advice 2 and 3 provide illustrative examples of the type of information that an ICM can provide to a human (corresponding to b) and c) in the above enumeration). That is, the confidence of the DSS is explained in terms of similar stored past experiences with its own performance. The difference between advice 2



The advice from KITT to drive manually or to turn on automatic mode is based on current information as obtained by sensors and a smart but complex algorithm. Although this advice is often correct, it may occur that it is not; sometimes the situation demands that you drive manually even though KITT advised you to turn on automatic mode, or vice versa.

Therefore, KITT tells you how confident it is that the given advice will prove to be correct. Below, we listed several ways in which KITT can express this confidence about the advise. We would like you to carefully read these expressions and rank them based upon your preference.

15. Please rank these five methods of how KITT can present its advice to you.

1 "I advise you to turn on automatic mode."

2 “I advise you to turn on automatic mode. I am very confident that I can handle this situation because I drove this road very often and did not require any manual override from you in 98% of these cases.”

3 “I advise you to turn on automatic mode. I am very confident that I can handle this situation because I drove this road on my own without any manual override except for one time. In this one situation, an oil stain was detected on a busy road and you chose not to evade it and switched to manual control.”

4 “I advise you to turn on automatic mode. I am very confident that I can handle this situation based on the current perceived road markings, weather conditions and amount of traffic and my knowledge about how I react to these.”

5 “I advise you to turn on automatic mode. I am very confident that I can handle this situation because in the future I expect no difficulties due to predicted weather and traffic conditions.”

Figure 5.7: Screenshot of the section of the survey in which participants were asked to rank different kinds of explanations based on their preference. Advice 2 and 3 provide illustrative examples of the type of information that an ICM can provide to a human.

and 3 is that the latter includes a specific example of a situation in which the advice appeared not to be correct, while the former does not.

5.6.4 Experimental design

We investigated two variables. 1) The importance of different information in determining when to follow an advice: information about the confidence measure's algorithm, information about prior experience, or information about the predicted future situation. 2) The information preference in an accompanying explanation: no additional explanation, general prior experience, specific prior experience, current situation, or predicted future situation. Both dependent

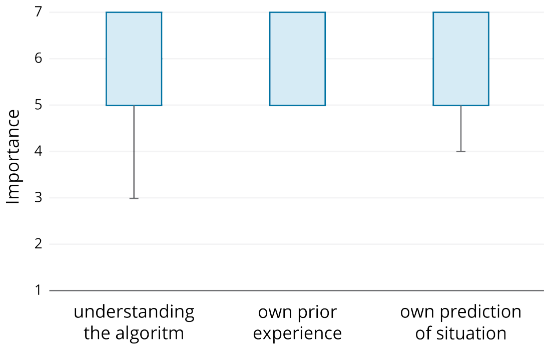


Figure 5.8: Boxplot of the Likert scale ratings indicating the importance of different types of information used to determine to follow the advice to turn on automatic driving mode. The higher the ratings, the more the information was preferred.

variables were investigated within-subjects, meaning that all participants indicated their importance rating and preference rankings for all types of information and explanations respectively.

5.6.5 Analyses

We performed two non-parametric Friedman tests with post-hoc Wilcoxon signed rank tests on the ordinal Likert scale data to investigate two topics: 1) The relative importance of information that taken into account when deciding whether or not to follow the advice, and 2) the difference between preference ratings of the types of explanation.

5.6.6 Results

Figure 5.8 shows the distribution of Likert scale ratings concerning the importance of information in the advice. Ratings are high in general, as indicated by the high medians and the minor deviations from these median scores.

There is a statistically significant difference in importance ratings of the considered information when evaluating an advice, $\chi^2(2) = 16.77$, $p < .001$. Wilcoxon signed-rank tests showed that participants rated prior experience with the system as more important for deciding about following an advice than understanding the advice system ($Z = -3.71$, $p < .001$), but not more important than predictions about future situational circumstances ($Z = -1.58$, $p = .115$). However, predictions about future circumstances were rated as being more important than understanding the advice system ($Z = -2.89$, $p = .004$).

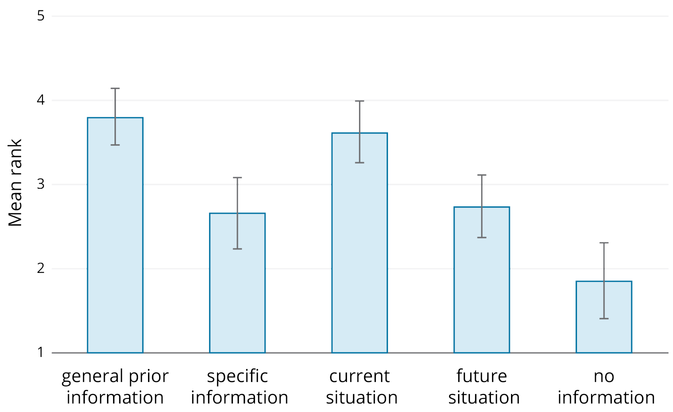


Figure 5.9: Means and 95% confidence intervals of the preference rankings concerning the different types of advice that are provided by the autonomous car. The rankings are inverted, the higher the rank the more preferred.

Figure 5.9 shows the means and 95% confidence intervals of the rankings concerning the preferences of participants for different types of additional information given in an advice.

There is a statistically significant difference in rankings of the five types of advice, $\chi^2(4) = 39.38$, $p < .001$. Table 5.5 shows the results of the post-hoc tests. Importantly, participants preferred the explanation that contained general prior experiences over the one that presented a specific experience of a case in which the advice was not followed. They also preferred general prior information over information concerning the future situation, and over no additional information. However, preference ratings for using general prior experience as explanation about an advice were, on average, not higher than using information about present situational circumstances.

In this human study, we investigated how participants judged various types of information a confidence measure may include in an explanation. Overall, the use of relevant prior experiences was judged as important in both defining confidence values and explaining them. Equally important was the information contained in the current situation. This indicates that ICM and its explanations match people's expectations and preferences of a confidence estimation. It underlines the importance of confidence values being explainable, something ICM readily supports. However, confidence measures may need to explain how the current situation relates to those past experiences. For ICM that entails explaining the similarity function and why it selected those past experiences. To do so, the similarity function needs to be easily understood or otherwise explainable.

	Present Situation	General past experience	Future situation	Specific past experience
Present situation				
General past experience	*n.s.*			
Future situation	$Z = -2.00$, $p = .045$	$Z = -2.35$, $p = .019$		
Specific past experience	$Z = -2.77$, $p = .006$	$Z = -3.34$, $p = .001$	*n.s.*	
No information	$Z = -3.86$, $p < .001$	$Z = -4.39$, $p < .001$	$Z = -3.40$, $p = .001$	$Z = -2.28$, $p = .023$

Table 5.5: Results of the Wilcoxon signed rank post-hoc tests on the preference rankings of information that is included in an explanation about the advice.

5.7 Discussion

Although the proposed ICM framework relies on a case-based reasoning approach, it is also closely related to the field of conformal prediction [231]. Methods from this field define a set of predictions that is guaranteed to contain the true prediction with a certain probability (e.g. 95%). Conformal prediction methods share many similarities to ICM, such as their model-agnostic approach and use of (dis)similarity functions. Current research focuses on making these methods more explainable and transparent [232]. Our experimental work on these topics may provide valuable insights for future conformal prediction methods. In addition, future work may aim to explore how conformal prediction methods can be used in the ICM framework.

An important trade-off in an ICM is between its accuracy and transparency, as an increase in accuracy implies an increase in complexity. A concrete example is the similarity measure, it can be as straightforward as Euclidean distance or as complex as a trained Deep Neural Network (as done by Mandelbaum et. al [223]). For some domains, a relative simple similarity measure may not suffice due to its high dimensional nature or less-than apparent relations between features (e.g. the many pixels in an image recognition task). A more complex or even learned similarity measure may solve such issues. However, it may may prevent humans from from adopting the system in their work due to a lack of understanding [88]. This is sometimes referred to as the accuracy and transparency trade-off in current AI. To solve this, simplified model-agnostic methods generating explanations may be a solution. However, it also requires exploring where humans allow for system complexity and where transparency is required.

Besides such technical issues, an interesting finding from the online survey was that participants did not found it important for an explanation to refer to past situations in which the provided advice proved to be incorrect. This could indicate the tendency of people to favor information that confirms their preexisting beliefs and to be ignorant towards falsification, a phenomenon known as the confirmation bias [233]. Importantly, such a preference does not necessarily mean that it is best to omit this kind of information. That is, the main goal of the transparency and explainability properties of an ICM is to enable humans to better understand where the confidence value originates from in order to more accurately predict the extent to which an advice of the system can be trusted. In order to enable people to make an accurate assessment, it is essential to provide both confirming and contradictory information, precisely because we know that people are prone to ignore information that does not confirm their beliefs. Future work on confidence measures should not only conduct human studies revolving around preferences, but also on how they affect system adoption, usage and task performance.

Moreover, findings from our human studies implied that people prefer to know about the current situational circumstances. This preference holds even when a given confidence value was high and they said they trusted this estimation. This could indicate that people still want to be able to form their own judgement about the DSS' advice based on their own observations, in order to maintain a sense of control and autonomy [234]. Hence, a confidence measure is not a substitute for a human's own judgement process and should be designed to facilitate this process. ICM's property of explainability may offer a vital contribution to this process. Further investigation is required to identify what should be explained in addition to an ICM and how this should be presented.

5.8 Conclusion

In this chapter we proposed the concept of Interpretable Confidence Measures (ICM). We used the idea of case-based reasoning to formalise such measures. In addition, we motivated the need for confidence measures to be not only accurate, but also explainable, transparent and predictable. An ICM aims to provide human collaborating with a decision support system (DSS) information whether the provided advice should be trusted or not. It does so by conveying how likely it is that the given advice turns out to be correct based on past experiences.

Three straightforward ICM implementations were proposed and evaluated, to serve as concrete examples of the proposed ICM framework. Two human studies showed that participants were able to understand the idea of case-based reasoning and that this was in line with their own reasoning about confidence. In addition, participants especially preferred their confidence values to be explained by referring to past experiences and by highlighting specific experiences in the process.

5.9 Acknowledgements

This research was funded by TNO's early research programs (ERP) and risk-bearing exploratory research programs (RVO). We would like to thank our colleagues for insightful discussions. Special gratitude go to Alexander Boudewijn, Stephan Raaijmakers and Catholijn Jonker. Finally, our gratitude to the dynamic positioning operators who found the time to participate in our experiment and share their experience with us.

CHAPTER 6

ACTIONABLE EXPLANATIONS FOR CONTESTABLE AI

This chapter is adapted from the paper under review; **van der Waa, J.**, van Diggelen, J., Neerincx, M. & Jonker. C. (2022). Actionable Explanations for contestable AI. *Journal of Artificial Intelligence Research*. The adaptations include an altered abstract, shortened introduction, and an adjusted lay-out, including the formatting of figures and tables.

Jasper van der Waa, Jurriaan van Diggelen, Catholijn Jonker, Mark Neerincx

While the previous chapter discussed the use of confidence and its explanations to calibrate one's trust. This chapter addresses the human's ability to contest the AI agent's decision and take action to alter it. As AI agents govern more parts of our lives, humans will find themselves subjected to their decisions, either directly (e.g., the decision involves the initial automated screening for a loan application) or indirectly (e.g., the decision initiates more scrutinized research into someone's finances). Such decisions will often prove to be unfavourable for our human goals (e.g., getting a loan or not to be registered as a fraud). Humans should have the ability to contest an AI agent's decision. Such an AI agent can support that ability by providing actionable explanations, making clear how its decisions can be altered through appropriate action. This chapter formally defines six properties that make an explanation actionable. This formalisation enables univocal comparisons and argumentations on explanation theories or models that contribute to contesting AI agents' decisions and provide the steppingstones for the development and testing of methods to generate such explanations. A literature review showed not all "actionable properties" are being addressed appropriately. Current explanations are faithful to the AI agent's functioning and of counterfactual nature, with an increase in attention to the explanation's interpretability. However, current actionable explanations do not convey explicit and feasible action suggestions that acknowledge human preferences. We conclude with a call to the research community to address these research gaps more actively such humans can maintain their autonomy even when subjected to decisions made by an AI agent.

6.1 Introduction

Within the field of Explainable AI (XAI) the main referred purpose of explanations is to calibrate trust by justifying an AI agent's decision or behaviour. See for instance Miller et. al [68], Hoffman et. al [61], Shin et. al [235], Neerincx et. al [236] and Waa et. al [237]. Trust calibration is of importance when the human and AI agent collaborate to make a best possible decision [236]. However, AI agents are increasingly used in cases where they make a decision about a human who is then subjected to that decision. Examples of such use cases are the automated processing of loan applications [238] or the initial filtering of job-applicants [239]. Even in use cases where human and AI agent collaborate in their decision, there is often another human subjected to that decision. For example in the case where a doctor is aided in their diagnosis of a patient with he help of an AI agent [240]. In these examples the life and autonomy of a human is affected by, in part, what an AI agent determines.

However, there is an intrinsic human value for humans to retain autonomy over their own lives, even when parts of their lives are governed by AI agents [241]. Autonomy, or self-determination, can be retained by enabling AI agents that are *contestable* by those who are subject to the AI agent's decisions. The field of XAI can help support this contestability of AI agents by providing explanations that are actionable [242], the term "algorithmic recourse" is also at times used to refer to contestability [29]. This implies that the explanation conveys the understanding humans need to identify the appropriate action needed to contest the AI agent's decision in a way such that this AI agent makes a more preferable decision about them. Such action might entail directly influencing the situation such that the AI agent makes a different decision [29]. However, it might also entail actions where humans are supported in filing sufficiently motivated complaints or requests to an appropriate oversight committee who can decide to recall and improve the AI agent [243]. The former is useful when the AI agent uses information that can be corrected or otherwise influenced by the human, whereas the latter is useful when the AI agent perceived as being unfair or malfunctioning.

To provide humans with the ability to effectively contest and alter an AI agent's decisions, they need an understanding about how such decisions are made and can be influenced through action [244, 179]. In other words, humans require actionable knowledge about an AI agent's internal decision making to contest its decisions. Not only could such actionable knowledge improve the collaboration and acceptance of AI agents, the need for contestability is emphasized in upcoming regulations proposed by the European Union to regulate the application of AI agents [245]. Although within the field of XAI there is increasingly more attention to support human contestability of AI agents through explanations, there is no consensus on what makes explanations support contestability. There is an overall lack in what defines an explanation as conveying actionable understanding to support contestability, explanations we refer to as *actionable explanations*. The lack of a clear definition of what makes an explanation actionable, limits us in determining whether such explanations actually support a human's contestability. This in turn prohibits the creation of a shared research agenda and the evaluation whether such explanations cause an AI agent to adhere to regulations requiring AI agent's to be contestable.

This work aims to remedy this and formally define six properties that would define an explanation as being actionable and thus supporting contestability (see Figure 6.1 for an overview). We do so by taking a socio-technical system (STS) perspective [246]. Such an perspective includes the AI agent, the human agent and their shared context. Three components are needed to address the challenge of contestability and to help formalize the notion of actionable explanations. As only by recognizing the effects an AI agent has on human agents and how such human agents can influence their shared situation can we fully address contestability.

We present a formal framework to define the role and purpose of an explanation generated and communicated by an AI agent. Next, this framework is used to formalize the six proposed properties. This formalization aims to remove ambiguity in their definition to foster scientific discussion, support the development of explanation generating methods, and the derivation of metrics to measure such properties. Finally, we perform a literature review of explanation generating methods whose explanations are referred to as supporting contestability. Each method is reviewed on whether it adheres to any of the six properties, with the aim to identify potential research gaps. This review shows the current state of the art on actionable explanations and open research challenges that still need to be resolved to truly support humans' contestability of AI agents.

Throughout this work we will illustrate our reasoning with a concrete example of an AI agent that functions as a vaccination planning tool during a pandemic. The tool's purpose is to plan vaccination dates given their medical records and lifestyle. Imagine someone receiving a date three months from now. That person however expected to be vaccinated within this month as it believes it should receive priority over others. Contestability in this example means that this person is capable of identifying the most effective action that would res-

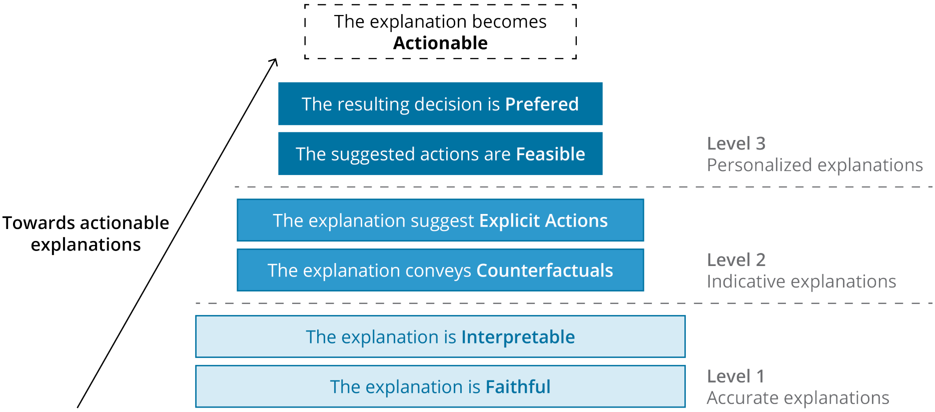

The explanation becomes **Actionable**
The resulting decision is **Prefered**
The suggested actions are **Feasible**
Level 3
Personalized explanations
Towards actionable explanations
The explanation suggest **Explicit Actions**
The explanation conveys **Counterfactuals**
Level 2
Indicative explanations
The explanation is **Interpretable**
The explanation is **Faithful**
Level 1
Accurate explanations

Figure 6.1: An overview of the proposed six properties that define and explanation as actionable. An actionable explanation is an explanation that provides humans with the understanding needed to effectively contest an AI agent's decision. These properties are separated into three levels of increasing complexity.

ult in a more favourable vaccination date or a tool more in line with their own values. We will use this example of a vaccination planning tool throughout the paper. See below for an actionable explanation that adheres to all six proposed properties:

> “Your vaccination is in three months due to your good health. Your records indicate you previously risked obesity. If this would still be the case with no other changes, your vaccination date would be in two weeks. If you believe you still risk obesity, contact your general physician to verify this and update your medical records accordingly, which would initiate a reschedule from three months to possibly two weeks."

This chapter is structured as follows. First, in Section 6.2 we introduce our socio-technical perspective towards contestability and actionable explanations. We then propose in Section 6.3 a formal framework to formalize explanations between the AI and human agent from such a perspective. This framework is then used to formalize of each of the six proposed properties, separated in three consecutive levels: an actionable explanation should be accurate (Section 6.4), indicative (Section 6.5), and personalized (Section 6.6). Next, Section 6.7 presents our literature review and identifies future research directions. Finally, we discuss and reflect on our work in Section 6.8 followed by our conclusions in Section 6.9.

6.2 Actionable explanations in a socio-technical system

See Figure 6.1 for an overview of the six proposed properties. They are subdivided into three levels. The first level states that the explanation should be accurate, which implies that the explanation should be faithful to the AI agent's reasoning and inner working as well as interpretable to the human agent. The second level describes that the explanation should be indicative. We will argue that this implies that the explanation should describe alternative situations in which different decisions will be made (e.g., counterfactuals) as well as suggest explicit actions to arrive at such alternative situations. Lastly, we will argue that this explanation should contain actions feasible for the human agent to perform as well as resulting in an AI agent's decision deemed preferable by that human agent. Thus, the third and final level describes the personalization of the explanation. Together, these properties define what would make an explanation actionable to support one's contestability.

The six properties are motivated and defined through a socio-technical system perspective. Even when a system is technologically sound, its functioning on deployment is not guaranteed [246]. This typically occurs when in the design the social and organisational context of its application is omitted. A socio-technical systems (STS) perspective tries to remedy this, as it ensures consideration of application context in system design [247]. It supports the reasoning about the effects a system brings about in the application context, whether such effects are preferred, and if not, how they should be addressed.

With the notion of contestability, the AI agent, human agent and their shared context come together. Contestability is the human's ability to effectively identify the actions needed to alter the shared context to induce a more favourable decision from the AI agent [242]. This ability is also referred to as algorithmic recourse [29]. Current research towards actionable explanations to support contestability focuses on conveying counterfactuals. A counterfactual explanation conveys alternative situations where the AI agent would make a different decision [29]. It is hypothesized that this enables human agents to select the alternative situation with a more favourable outcome and infer which actions to take to arrive at that situation. Although the communication of counterfactuals recognizes the relation of the AI agent with its own situational context, they forego the human agent and its context, which risks resulting in sub-optimal actionable explanations. For example when the human agent is incapable of correctly interpreting the counterfactuals and incapable of inferring an effective contesting action from them.

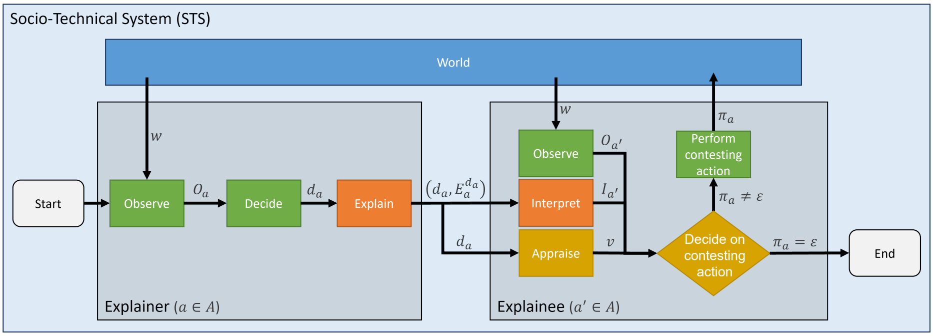


Figure 6.2: A socio-technical system (STS) perspective on how explanations can support the explainee to take appropriate action when an AI agent's decision is deemed unfavourable.

We argue that an STS perspective towards actionable explanation results in better explanations to support contestability. Specifically that the recognition of the AI and human agent as well as their shared context allows the field of XAI to design more effective actionable explanations. In Figure 6.2 we illustrate the STS focused on contestability and the role of explanations therein.

This figure shows the AI agent, referred to as the explainer, and the human agent, referred to as the explainee, interacting with each other in a shared world. Both the explainer and explainee *observe* this world, inferring their own unique set of observations. The explainer uses these observations to *decide* on a decision or advice which it attempts to *explain*. The explainer communicates both to the explainee. Who *interprets* the explanation, adding to its understanding how the explainer makes decisions, and *appraises* the decision, reflecting how pleased the explainee is with the decision. The explainee combines this interpretation, appraisal and its observations of the world to determine whether action is required, useful or possible to contest the explainer's decision. This is the process where the explainee *decides on a contesting action* if required and which action that should be chosen to alter the explainer's decision into a more favourable one. When such an action is determined, the explainee *performs* it, changing the world in a way that ideally results in the intended effects and a more favourable decision.

In our running example the vaccination planner tool is the explainer, and the person querying it for a vaccination date the explainee. The process of observing the world means for both that they know the explainee's medical records and lifestyle. The vaccination planner tool, uses these observations to plan a date three months from now, a decision that affects the explainee's life. The explainee subsequently appraises the decided date and feels this is unfavourable. If no explanation would be given, the explainee can only rely on its own experience, knowledge and assumptions to decide upon an action to try and change this decision. However, the tool does provide an explanation which the explainee can interpret. This explanation becomes actionable when it directly supports the explainee to determine which action to take to contest the explainer's decision.

In the explanation example given in the introduction, the planner tool justifies its decision by referring to the observation that the explainee is in good health. The proposed properties *faithfulness* and *interpretable* state that this justification adheres to the tool's reasoning and is interpretable by the explainee. The explanation is also a *counterfactual* explanation, the third property. In this example the communicated counterfactual is the alternative context when the explainee would still suffer from a risk for obesity. This would cause the vaccination date to fall within two weeks instead of three months. The explanation also contains *explicit action* suggestions, thus adhering to the fourth property. It suggests the action to contact its physician to report the risk of obesity, which could result in an updated medical record causing a rescheduling of the vaccination date. Finally, the explanation is *feasible* and *preferable* since the explainee is able to contact its physician and would favour a vaccination date within two weeks over three months. These final two properties limit the number of relevant counterfactuals to be communicated to only those which the explainee can act upon.

In the next sections we formally define each of these six properties to explicate their meaning and remove any ambiguity. We do so given a formal framework based on the socio-technical system perspective outlined in Figure 6.2. The next section provides this framework.

6.3 Formal framework to explanations

Below we describe the framework used to formalize the proposed properties. Below we introduce we list the sets used in our formalization.

- **Worlds.** W is the set of possible worlds, where $w, w' \in W$ are variables ranging over worlds. For example, two worlds might differ in the person querying the planner tool with different medical records and lifestyles.

- **Agents.** A is the finite set of agents, where $a, a' \in A$ are variables ranging over agents. For example the vaccination planner AI and the human querying it could be agents in A.

- **Decisions.** D is the finite set of all decisions that agents can make. Where $D_a \subseteq D$ is the finite set of decisions agent a can take. Where $d \in D$ is a variable ranging over D, and similarly $d_a \in D_a$ a variable ranging over the possible decisions from agent a. For example, the decision of the scheduled date for vaccination by the vaccination planner agent.

- **Contesting actions.** Π is the finite set of all possible actions agents can undertake in the world to contest an agent's decision. With $\Pi_{a'} \subseteq \Pi$ the set of all actions agent $a' \in A$ can undertake. Where $\pi_{a'} \in \Pi_{a'}$ is a variable ranging over single actions from a'. For example, to contact one's physician to update the used medical records. We write $\epsilon \in \Pi_{a'}$ to signify taking no action.

- **Decision-making functions.** Q is the set of decision-making functions that describe the decision-making processes of agents. Let $Q_a \subseteq Q$ denote the set of decision-making functions of agent a. Where $q \in Q$ is a variable ranging over Q, and similarly $q_a \in Q_a$ a variable ranging over all decision-making functions of agent a. For example the rule to prioritize people with a risk for obesity for vaccination. Similarly, a counterfactual is also such a function, as it is a specific set of observations (medical records) with its associated decision (vaccination in three months).

  For any $Q' \subseteq Q$ we use $\models_{Q'}$ to express that a decision making process satisfies the decision making functions of Q'. Formally, we define this recursively as follows.

  The base case is: $\forall q, w, d_a$: $w \models_{\{q\}} d_a$ *iff* $d_a \in q(w)$. If no confusion is possible, we write $\models_q$ instead of $\models_{\{q\}}$. The general case is: $\forall Q' \subseteq Q$, $\forall w, d_a$: $w \models_{Q'} d$ *iff* $\exists q \in Q' : w \models_q d$.

  In general, we assume that for any agent a, the set Q_a is consistent. Consistency means that in the same state there are no decision functions applicable that would lead the agent to make decisions that contradict each other.

- **Explanations.** $\mathcal{E}$ is the set of possible explanations. With $\mathcal{E} = 2^Q$ which denotes that explanations consist of one or more decision-making functions. Where $E_a^{d_a} \in \mathcal{E}$ is the explanation from agent a about a decision d_a. For example, $E_a^{d_a}$ can contain the decision rule that people risking obesity are prioritized for vaccination, when explaining the decision to plan the vaccination in three months.

These sets will be used to formalize each of the distinct steps shown in Figure 6.2. We do so following a multi-agent system formalization that makes use of the notion of beliefs. Beliefs represent (possibly) imperfect information which are believed to be true by some agent [248]. For instance, an agent can have beliefs about the world state (e.g., observations [248]), its own decision making (e.g., introspective beliefs [249]), and of other agents (e.g., mental models [250]). This offers a basis to formally define explanations and their properties, as explanations can be viewed as formulating, conveying and interpreting beliefs between agents for a certain purpose.

The notion of beliefs is used in combination with the above sets to formalize the processes from Figure 6.2. We denote $Bel_a(\,\cdot)$ as agent $a \in A$ believing that a statement · is true. For example, $Bel_{a'}(\,q \in Q_a)$ signifies that a' believes that a makes decisions according to decision-making function q. This q can be communicated in an explanation. We can then use this to formalize the interpretation of an explanation.

Below we formally define the processes from Figure 6.2, starting with those of the explainer.

6.3.1 Explainer processes

The *explainer* $a \in A$ from Figure 6.2 makes use of three processes; observe the world, decide on a decision or advice, and explaining that decision to an explainee. Below we list the definition of each, making use of the previously discussed sets.

Definition 6.3.1 (Observe)**.**
Let $a \in A$ be an agent, then $O_a : W \longmapsto W$ is the agent's observation function. If w is the current world state, $O_a(w)$ denotes the agent's current observations. The agent believes what he observes, i.e., $Bel_a(\,O_a(w))$. Ideally, the observations are correct, such that $O_a(w) \subseteq w$. We write O_a to denote the current observations by the agent if no confusion is possible.

The planner tool from our example observes the medical records of anyone querying it. Those records are part of the world, and are the observations the planner tool agent makes about it. Note that this process is also occurs for the explainee, $a' \in A$, whose current observations are denoted as $O_{a'}$.

Definition 6.3.2 (Decide)**.**
Let $a \in A$ be an agent, then $O_a \models_{Q_a} d_a$ is its decision-making process based on its own current observations and decision-making functions, with $d_a \in D_a$ as the made decision.

With our running example, the decision for vaccination in three months or two weeks are decisions. The planner tool makes these decisions based on its observed medical records and according to its set of decision making functions, which can be anything, e.g. a set of rules or a deep neural network.

Definition 6.3.3 (Explain)**.**
Explaining leads to an explanation $E_a^{d_a} \in \mathcal{E}$ about the decision d_a provided by agent a. We assume that agent a believes its explanations in the sense that it believes that the decision d_a is made according to $E_a^{d_a}$. Formally, $Bel_a(O_a \models_{E_a^{d_a}} d_a)$.

Recall that explanations consist of one or more decision-making functions ($\mathcal{E} = 2^Q$). In our example the explanation consists out of two of such functions. First, the planner tool conveys that people risking obesity are prioritized (a decision-rule, which is a decision-making function). Secondly, if nothing changed except for the querying person to have that risk, its vaccination would be in two weeks instead of three (a counterfactual, a very specific decision-making function). Note, that the planner tool's actual decisions do not necessarily have to be made according to these two. For instance when the tool's explanation or decision process is faulty. We only assume that the planner tools believes it makes decisions according to them.

6.3.2 Explainee processes

The *explainee* $a' \in A$ from Figure 6.2 makes use of four unique processes. That of interpreting an explanation, appraising the decision from the explainer, deciding if action is needed, and perform that action if so. We define these as follows:

Definition 6.3.4 (Interpret)**.**
The interpretation of agent a' for some explanation is done by the function

$I_{a'} : D \times \mathcal{E} \longmapsto \mathcal{E}$. Given an $E_a^{d_a}$ from agent a about decision d_a, the interpretation $I_{a'}(E_a^{d_a}, d_a)$ are all decision-making functions of which agent a' believes that a uses them to make its decision d_a; $Bel_{a'}(\, I_{a'}^{d_a} \subseteq Q_a)$. We abbreviate $I_{a'}(E_a^{d_a}, d_a)$ to $I_{a'}^{d_a}$.

Ideally, $I_{a'}^{d_a} = E_a^{d_a}$, but, in general, this need not be the case. In the running example, the explanation conveys the decision rule that people risking obesity are prioritised. However, the person querying the planner tool (the explainee), might have a different interpretation of what constitutes as "risking obesity" than the planner tool (the explainer). For instance, the interpretation could become that people weighing more than the explainee receive priority, which is not necessarily a correct interpretation of the explanation.

Definition 6.3.5 (Appraise).
Appraising a decision leads to the appraisal $v \in \mathbb{R}$ of a decision d_a by the explainee a'. Let v' be the appraisal for any $d'_a \in D_a$ where $d'_a \neq d_a$. Then $v > v'$ denotes that d_a is favoured over d'_a and $v \leq v'$ denotes d_a is favoured less than d'_a by agent a'. The explainee's appraisal thus defines D_a as a total ordered set unique to that explainee.

The above definition relies on the fact that people have a certain preference for decisions and that they are (dis)pleased with a decision when a certain threshold is reached. In the example, this could mean that the explainee would favour any vaccination date within 2 months equally well, but any date later then that would be less and less preferable.

Definition 6.3.6 (Decide on contesting action).
The decision for an contesting action leads to a given action $\pi_{a'} \in \Pi_{a'}$ by a' based on the appraisal v of the current decision d_a. Where $\pi_{a'} \neq \epsilon$ iff $v < t$, with $t \in \mathbb{R}$ being an arbitrary threshold, meaning that the explainee decides on an action when the current decision is deemed unfavourable. We assume that $\pi_{a'}$ is made on the basis of its interpretation $I_{a'}^{d_a}$ and observations $O_{a'}$.

Combined with the appraisal, this definition follows the notion that at some point a decision is unfavourable enough to warrant action. Recall that ϵ was the "empty" action, signifying a decision to take no action. In the example this might have been the case if the vaccination date was within two months. Instead it is planned in three months, thus triggering the process of deciding on an actual action, which it will try to do based on its appraisal of the decision (e.g., reflecting its motivation), the interpretation of the explanation (e.g., if risking obesity the date would be in two weeks), and its own observations (e.g., currently risking obesity). Potentially deciding to indeed contact its physician as the explanation suggests.

Definition 6.3.7 (Perform contesting action)**.**
By performing a contesting action $\pi_{a'}$, a new world w' is achieved following from the current world w. This is expressed with the transition function $w' = Act(\pi_{a'}, w)$.

If the person querying our example planning tool decided to contact its physician, a world might be achieved where that person indeed runs the risk for obesity according to its medical records.

In this treatise, we assume that the human agent, the explainee, decides to take action based on its own observations, its appraisal of the decision and its interpretation of the explanation. Under this assumption, it is thus vital that the offered explanation supports the explainee in taking the correct action. When it offers this support, that explanation is referred to as an actionable explanation.

Below we formally define each of our six proposed properties that would make an explanation actionable. We make use of the above defined sets and processes of both explainer and explainee, which we typically denote as a and a' respectively.

6.4 Level 1: Accuracy

The first two properties for an actionable explanation revolve around the explanation's accuracy. To best support contestability, an explanation needs firstmost to be accurate, hence we refer to this as a Level 1 actionable explanation. However, explanation accuracy is an ambiguous concept with many different interpretations and perspectives [32]. To remove this ambiguity we distinguish between two aspects of accuracy, namely *faithfulness* and *interpretability*. The former addresses how sound the explanation is to the explainer's decision-making. The latter addresses how resembling the interpretation of the explainee is to the explanation. We argue that both are required to identify an explanation as accurate.

In theory, when an explanation is both faithful and interpretable, the explainee could rely on that information to decide if and how to act. However, the explainee is limited in estimating the faithfulness and interpretability of an explanation. As it is unlikely that the explainee understands the explainer sufficiently to identify an explanation as unfaithful. It is even less likely that the explainee would be aware of an incorrect interpretation on its own end. As such, assuring that explanations are faithful and interpretable is the responsibility of the explanations' designer.

6.4.1 Explanation faithfulness

Informally, a faithful explanation is an explanation whose conveyed description of a decision-making process is correct compared to that same process: the explanation contains no falsehoods. If an explanation is not faithful, it cannot be relied upon to incite a good understanding nor the inference of an appropriate action, which reduces the ability for the explainee to contest the explainer's decision based on this explanation.

If in our example, the planner tool would actually make its decision based on age instead of the risk for obesity, the communicated explanation would not be faithful. The supported contestability is reduced, as even when the explainee decides to contact its physician and it is decided that the explainee indeed risks obesity, nothing would change. Thus it is vital that explanations are faithful due to an explainee's dependency on them.

We formalize the faithful property as follows:

Property 6.4.1 (Faithful)**.**

Let $E_a^{d_a}$ be an explanation from explainer a about some decision d_a. Then *Faithful*$(E_a^{d_a})$ iff $O_a \models_{E_a^{d_a}} d_a$, where O_a is the current set of observations made by the explainer.

In other words, an explanation is *Faithful* when all the descriptive decision-making functions that the explanation contains leads to the explainer's decision. Given that the explainer only adds decision-making functions to its explanations if it believes that these apply, this definition implies that in case of a faithful explanation these beliefs are indeed true.

6.4.2 Explanation interpretability

Even a faithful explanation can be misinterpreted, resulting in incorrect inferences from that explanation by the explainee. When inferring which action to take to contest the explainer's decision, the supported contestability is reduced. Informally we define an explanation to be interpretable when its interpretation by the explainee results in exactly the same information as was conveyed by the explainer.

In the example, an incorrect interpretation would be to infer that all who weigh more than the person querying the planning tool have priority. As this does not necessarily follow a medical definition of risking obesity. This might incite a different kind of action, one based on a feeling of unfairness. Hence, it does not matter how faithful an explanation is, if its interpretation is not sound, the support to the explainer's contestability is limited.

We define an interpretable explanation as follows:

Property 6.4.2 (Interpretable)**.**

Let $E_a^{d_a}$ be an explanation from explainer a about decision d_a. Then *Interpretable*$(E_a^{d_a})$ iff $I_{a'}^{d_a} \equiv E_a^{d_a}$. [9]

Note that this definition does not explicitly require that the explainee has the exact interpretation of what was conveyed. Instead, an equivalent interpretation suffices. For instance, the interpretation "people with a weight above mine are prioritized" might be a correct interpretation of "people running risk for obesity are prioritized", if that person runs the risk themselves which would make the two decision-rules equivalent. Furthermore, as long as the interpretation is equivalent, the medium with which the explanation is conveyed does not matter. Whether visualized or textual, the above definition supports that both can result in the same interpretation.

To summarize, ideally we would like explanations to be both *faithful* and *interpretable*. This ensures that the knowledge that is communicated can reliably be used to infer any contesting actions.

[9] $\equiv$ means here that $\forall w \in W, \forall d \in D : w \models_{E_a^{d_a}} d \Leftrightarrow w \models_{I_{a'}^{d_a}} d.$

6.5 Level 2: Indicative

Besides being accurate, for an explanation to be actionable, it should specifically aim to support the explainee's ability to take action. Thus, we argue that actionable explanations should contain *counterfactuals* and be *explicit* in the contesting actions to achieve those. The former ensures that alternative observation sets and their subsequent decision are communicated, which help the explainee infer favourable decisions and when they are given. The latter ensures that the explanation also contains actions that result in aforementioned counterfactuals. This removes an additional inference step for the explainee.

These two properties create a more indicative explanation. Such an explanation better indicates what options are available to the explainee and how to achieve them, improving the support to the explainer's ability to contest decisions. Albeit, these proposed counterfactuals and actions should be faithful as well as correctly interpreted. Therefore, we refer to this as the second level of actionable explanation. It builds on the former level of requiring accurate explanations.

6.5.1 Counterfactual

The purpose of an explanation is to explain why a decision was made. For example through providing decision rules or behavioural examples. Counterfactuals are such example-based explanations. A counterfactual consists of 1) hypothetical observations different from the current observations and 2) the decision the explainer believes it would make in those hypothetical situations. Counterfactual explanations are viewed as a way to support contestability [29]. In a sense, a counterfactual tells the explainee; "if this and that would be observed, I would make this decision instead". Only the differences between the hypothetical observations and current observations tend to be explained [29]. This is done to minimize the explained information to only that what is vital [42].

We can formalize a counterfactual using our framework. We denote the hypothetical observations as O'_a. Then, using the decision-making functions the explainer follows Q_a, we can write $Bel_a(\, O'_a \models_{Q_a} d'_a)$. This expresses that the explainer believes it would make decision d'_a when observing O_a through the use of Q_a. We denote $(O'_a \rightarrow d'_a) \in W \times D_a$ as the counterfactual; the observations O'_a of which the explainer believes would lead to some decision d'_a.

The counterfactual $(O'_a \rightarrow d'_a)$ is a specific type of decision-making function. As it explicitly describes which decision will be made with what observations. Thus we can say that $(O'_a \rightarrow d'_a) \in Q$, and that $Bel_a(\, (O'_a \rightarrow d'_a) \in Q_a)$.

For an explanation to support contestability, we thus want it to contain counterfactuals. We define this as the property *Counterfactual*, denoting that the explanation contains counterfactuals. Formally;

Property 6.5.1 (Counterfactual)**.**

Let $CF_a(d_a) = \{(O'_a \rightarrow d'_a) \in W \times D_a \mid Bel_a(O'_a \models_{Q_a} d'_a) \wedge d'_a \neq d_a\}$, which represents the set of all possible counterfactuals whose decisions are different then the current decision d_a.

Then *Counterfactual*$(E_a^{d_a})$ iff $CF_a(d_a) \cap E_a^{d_a} \neq \emptyset$.

This defines any explanation as being *Counterfactual* when it contains at least one counterfactual from $CF_a(d_a)$, the set of all counterfactuals with a different decision. This property already limits the counterfactuals to only those with a different decision than the current one, with next properties limiting this set further. In our running example, the explanation contains one of such counterfactuals although only the actual differences with the current observations are communicated. Namely, it contains the counterfactual where the explainee would have the same medical records except for an added risk for obesity. With those observations, the planner tool would plan the vaccination in two weeks instead of three months.

The difficulty with this property is that so far it only limits the counterfactuals in an explanation based on them having a different decision than the current one. Besides that, there is no limit to the number of counterfactuals in the explanation. The properties in the next section will remedy this, for now the *Counterfactual* property simply states that an explanation should contain counterfactuals whose decision is different.

These counterfactuals enable the explainee to appraise each different decision, and select one decision that is more acceptable. In this case, when deciding on an action, the explainee only needs to infer which actions would lead to the explainer making those observations. In our example this would be the action to visit a physician and let that physician note a potential risk of obesity in the explainee's records. As such, the *Counterfactual* property is a step towards an actionable explanation that supports contestability.

This property relies on the properties *Faithful* and *Interpretable*. When the explanation is *Faithful* it means that a counterfactual $(O'_a \rightarrow d'_a) \in E_a^{d_a}$ is sound, meaning that the explainer indeed makes the decision d'_a when observing O'_a. It relies on the *Interpretable* property, in the sense that the explainee should understand both the observations and different decision from a counterfactual $(O'_a \rightarrow d'_a) \in E_a^{d_a}$. Otherwise the inference for a contesting action might be invalid. For example, if the explainee would not understand what it means to have "a risk for obesity", it cannot infer whether it has that risk but it simply is missing from the medical records.

6.5.2 Action explicitness

Even if the explanation contains counterfactuals, the explainee still has to infer suitable contesting actions that result in the explainer making those other observations. This can be difficult if the explainee is not a domain expert or may not understand how the explainer observes the world. Thus we argue that the explainer should *explicitly* convey action suggestions that could lead to a communicated counterfactual. Such suggestions simplify the explainee's inference for a suitable action.

We refer to an explanation conveying contesting action suggestions as being *Explicit*. Interestingly such suggestions do not explain anything about the explainer's decision-making, instead they explain how to achieve a more desirable decision from that explainer. The result is that the *Explicit* property implies an extension of the definition of an explanation. Where an explanation $E_a^{d_a}$ was defined as only containing decision-making functions, $E_a^{d_a} \subseteq Q$, we now extend this to also include actions. Any explanation who thus wants to support an explainee's contestability should contain more than just an explanation about how decisions are made. It should also explicitly suggest actions the explainee might want to take to achieve a different decision.

To formally define the property *Explicit*, we first have to define the process of selecting one or more actions as suggestions related to a counterfactual $(O'_a \rightarrow d'_a)$. We refer to this process or function as ρ, and we define it as follows:

Definition 6.5.1 (Action suggestion identification).
Let the function $\rho : CF_a(d_a) \mapsto 2^{\Pi}$ determine the set of contesting actions for counterfactuals, defined by $\rho((O'_a \rightarrow d'_a)) = \{\pi \in \Pi \mid Bel_a(O'_a \subseteq \mathit{Act}_{a'}(\pi, w))\}$.

Less formally, this function ρ takes a counterfactual $(O'_a \rightarrow d'_a)$ and selects one or more contesting actions. The explainer believes that when any of these actions is performed by the explainee the world changes in a way that causes the explainer to observe O'_a instead. In our example, this function ρ identifies the action to contact a physician to achieve a change, a risk for obesity, in the explainee's medical records, which is the action related to the counterfactual where with a risk for obesity the vaccination is planned in two weeks.

We can use this function ρ to formalize the *Explicit* property. Specifically, we can use it to obtain and add actions suggestions for every counterfactual in the explanation. More formally:

Property 6.5.2 (Explicit)**.**

Let $E_a^{d_a}$ be a counterfactual explanation from explainer a about decision d_a.

Let the function $\mathrm{X} : \mathcal{E} \mapsto \mathcal{E} \bigcup (\mathcal{E} \times \Pi)$ be defined as follows:

$$\begin{aligned}\mathrm{X}(E_a^{d_a}) = &\left\{ q \in E_a^{d_a} \mid q \notin CF_a(d_a) \right\} \bigcup \\ &\Big\{ \langle \rho\left((O'_a \to d'_a)\right), (O'_a \to d'_a) \rangle \mid \\ &\quad (O'_a \to d'_a) \in E_a^{d_a} \wedge (O'_a \to d'_a) \in CF_a(d_a) \Big\}\end{aligned}$$

This function adds action suggestions to every counterfactual in the explanation.

We call $\mathrm{X}(E_a^{d_a})$ an *Explicit* explanation.

This function X takes an explanation and make all of its counterfactuals explicit by applying ρ to get the suggested contesting actions. These actions are combined with the counterfactual. So for an arbitrary counterfactual $(O'_a \to d'_a) \in E_a^{d_a}$, the function X extends it to $\langle \pi_{a'}, (O'_a \to d'_a) \rangle$ if $\{\pi_{a'}\} = \rho((O'_a \to d'_a))$. In our running example, $\langle \pi_{a'}, (O'_a \to d'_a) \rangle$ is the counterfactual with a date in two weeks when a risk of obesity is in the medical records, and $\pi_{a'}$ is the suggestion to contact a physician to check and add such a risk to the records.

When an explanation is *Explicit* it specifically supports the explainee's ability to contest decisions through actions, thus it becomes a more actionable explanation. It reduces the amount of inference the explainee has to perform to decide upon a contesting action, as it can review each suggestion and decide whether it is worth it.

6.6 Level 3: Personalized

An issue with both a *Counterfactual* and *Explicit* explanation is the large number of possible counterfactuals and action suggestions. A selection needs to be made which will prove to be the most effective in supporting selecting the most appropriate and effective action. The communicated action suggestions should be limited to those that are *Feasible* for the explainee to perform. Whereas the communicated counterfactuals should be limited to the *Preferable* alternative decisions the explainee. We argue that these two properties lead to a natural way of reducing the number of communicated counterfactuals and action suggestions to those that effectively support contestability.

The properties *Feasible* and *Preferable* lead to a personalized explanation. Ideally, with a single counterfactual that leads to the most preferred alternative decision, combined with action suggestions the explainee is capable of performing.

6.6.1 Feasible

In particular, the idea of a feasible explanation is not novel in the literature on creating explanations supporting contestability [251]. This property is often implemented as the similarity between the current observations O_a and the alternative observations O'_a from $(O'_a \rightarrow d'_a)$. The greater the similarity, the more feasible the counterfactual is assumed for an explainee to accomplish. However, this notion is being critiqued as observational differences do not need to be in proportion to effort [252]. Some observations might be changed relatively easy by the explainee (e.g., correcting a mistake in the medical records), others take more effort (e.g., adjusting one's lifestyle) and others can be impossible (e.g., changing one's age). We agree with these criticisms; that the feasibility of achieving an counterfactual is the degree to which the explainee is capable of performing the required actions, not the differences those actions effectuate in the explainer's observations.

We thus define a *Feasible* explanations, as an explanation containing action suggestions that are part of the set of contesting actions the explainee can take. To formally define this, we introduce a function Γ which identifies all such action suggestions in an (explicit) explanation. If all these actions are part of the explainee's actions $\Pi_{a'}$, the explanation is deemed feasible. More formally:

Property 6.6.1 (Feasible)**.**

Let $\mathrm{X}(E_a^{d_a})$ be an explicit explanation from the explainer a about some decision d_a.

We define the function $\Gamma : \mathcal{E} \bigcup (\mathcal{E} \times \Pi) \mapsto \Pi$ inductively as follows:

- $\Gamma(\emptyset) = \emptyset$

- $\Gamma\Big(\langle(O'_a \rightarrow d'_a), \rho((O'_a \rightarrow d'_a))\rangle\Big) = \rho((O'_a \rightarrow d'_a))$

- $\Gamma\Big(\mathrm{X}(E_a^{d_a})\Big) = \Big\{\rho((O'_a \rightarrow d'_a)) \mid \langle\rho((O'_a \rightarrow d'_a)), (O'_a \rightarrow d'_a)\rangle \in \mathrm{X}(E_a^{d_a})\Big\}$

This function identifies and extracts all contesting actions within an explicit explanation.

Then *Feasible*$(\mathrm{X}(E_a^{d_a}))$ iff $\Gamma(\mathrm{X}(E_a^{d_a})) \subseteq \Pi_{a'}$.

This property dictates that for an explicit explanation to be feasible, all its proposed contesting actions should be possible to perform by the explainee. In our running example the explainee is deemed capable of contacting its physician to update its medical records with a risk for obesity if applicable. Another suggested action could have been to adjust the explainee's lifestyle such that this risk of obesity is guaranteed. Both actions might have the same result; a risk for obesity, causing a prioritizes the explainee's vaccination. However, the latter action is not part of the set of actions the explainee as it already believes it has a risk for obesity and is not willing to change it lifestyle to increase it.

Previously we argued that an explicit explanation improves contestability. With the given explicit action suggestions an explainee only has to choose instead of also inferring them. When these action suggestions are also all guaranteed to be *Feasible*, contestability is further improved, as the number of suggestions are reduced to those that matter for the explainee.

6.6.2 Personalized content

The above definition of *Feasibility* limits the action suggestions to those that the explainee is capable of. However, they do not limit the number of counterfactuals that could be communicated. We argue that the natural way of doing so is to account for what decisions the explainee prefers the explainer to make. Thus only communicating counterfactuals in its explanation that result in an alternative decision that is more preferred then the current one. Ideally, we would like this to be the most preferred decision the explainer can take. In other words, we want the explanation to communicate the *Preferable*, which offers a way to intelligibly select counterfactuals while further supporting the explainee's contestability.

For an explanation to become *Preferable*, all of its counterfactuals should have alternative decisions that appraised more than the currently made decision. Formally:

Property 6.6.2 (Preferable)**.**

Let *Counterfactual*$(E_a^{d_a})$ be an explanation from the explainer a about some decision d_a. Also, let v be the appraisal of d_a by a'.

Then *Preferable*$(E_a^{d_a})$ iff $\forall(O'_a \rightarrow d'_a) \in E_a^{d_a} : v' > v$ where v' is the appraisal of d'_a.

According to this definition, an explanation is deemed preferable if all its counterfactuals result in a better appraisal then the currently made decision, which implies that the explainer is aware of what the explainee prefers in decisions. In our running example, this means that the vaccination planner tool is aware that a vaccination date within two weeks is preferred more than a date within three months.

The property *Preferable* offers a way to select counterfactuals from the potentially infinite possible counterfactual. In doing so, it also improves the explainee's contestability as the explanation only addresses what needs to change to get a more preferred decision.

6.6.3 Actionable explanations

To summarize, we introduced three levels of explanation properties; 1) accurate explanations, 2) indicative explanations and 3) personalized explanations. The first level states that explanations should be faithful and interpretable. The second level states that the explanation should contain counterfactuals as well as explicit action suggestions. Finally, the third level states that these counterfactuals should be limited to those with decisions the explainee prefers while the action suggestions are feasible for the explainee to perform.

For an explanation to support the explainee's ability to contest the explainer's decisions, all three levels need to apply to that explanation. We define such an explanation as *Actionable*. As it not only explains how the explainer makes decisions, but does so in a way that supports the explainee to contest the explainer's decisions. When explanations are accurate, their content can be relied upon as well as understood. When explanations are indicative, they reduce the explainee's required inference based on knowledge they might not have. Finally, when explanations are personalized, they recognize the potentially unique needs of the explainee. Thus conveying reliable and interpretable knowledge that allows the explainee to take control over their life, even when that life is partially governed by decisions made by AI agents.

6.7 Future research areas for actionable explanations

A concise literature review was conducted to map the state of the art explanation generating methods to the above defined properties. The aim of this review is to find the current state of achieving actionable explanations and future research areas.

The review was conducted by combining several literature survey papers on explanation generating methods. These were the works from Adadi et. al [38], Guidotti et. al [39] and Arrieta et. al [20]. This resulted in a total of 237 papers on methods. These were then pruned to only those methods that stated to address the topic of contestability. This resulted in a total of 75 unique methods who aim to generate an explanation to help the human to contest the AI agent's decisions.

In Table 6.1 we show the percentages of the 75 reviewed methods per addressed, mentioned or not present property. Table 6.2 shows the complete overview for each method. No method seems to currently adhere to all six properties.

Most of the reviewed methods (81%) aim to identify or generate counterfactuals from a given set of alternative situations. These methods can often be led back to the approach proposed by Wachter et. al [29]. They propose the use of a loss or optimisation function that assesses how well a given counterfactual suits some measurable aspects of a counterfactual. Originally the only two aspects accounted for were whether the counterfactual caused a different decision and the proximity of that counterfactual to the current situation. The former could be viewed as the faithfulness of the explanation and most of the reviewed papers perform such an analysis (59%). However, the latter only roughly approximates the idea behind our property of feasibility. Our definition of feasibility applies to how feasible the explicitly proposed actions are to perform by the human explainee. Since most methods do not propose actions to achieve their identified counterfactual, they cannot adhere to the property of feasibility.

In fact, only very few of the reviewed methods address the property of generating explicit explanations (4% does). All of these methods introduce the notion of linking particular actions to a change in a situational attribute. In particular the method by Ramakrishnan et. al [56] stands out. They describe the idea to link actions to situational changes and take this a step further by also assigning weights to such actions. This allows their method to personalize explanation towards what is feasible for the human to perform. A lower weight would directly model a more feasible action and allow feasibility to be modelled in the loss or optimisation function used so commonly to identify counterfactuals.

The most underrepresented property in the reviewed methods (3%) is that of using counterfactuals that result only in a preferable decision for that particular human. Interestingly it is the method proposed by Wachter et. al [29] that attempts to adhere to this property. Although their work formed the basis of most of the reviewed methods, none of the reviewed methods continued with this aspect of their work. They argued that the loss or optimisation function used to identify counterfactuals, should be tailored to include what the human deems as a more favourable decisions. The other reviewed methods only required the decision to be different, not necessarily favourable (or their exemplar use case only involved two potential decisions).

Finally, the property of generating interpretable explanations is not often addressed in the reviewed methods (only 16% did so). Those that did, performed a limited user study to assess this property. Limited in the sense that they introduced participants with a proxy task and relied upon subjective measurements only. An approach for which Doshi-Velez and Kim [33] argue that it mostly evaluates a method's face validity but does not contribute to the field of XAI with a more theoretical insight why certain explanations are more interpretable than others. Nonetheless, these works did illustrate the interpretability of their pro-

	Level 1: Accurate explanations		Level 2: Indicative explanations		Level 3: Personalized explanations	
	Faithful	Interpretable	Counterfactual	Explicit	Feasible	Preferable
Addressed	**59%**	16%	**81%**	4%	7%	3%
Mentioned	28%	31%	5%	8%	29%	11%
Missing	13%	**53%**	13%	**88%**	**64%**	**87%**

Table 6.1: The percentage of the reviewed papers (75 in total) on explanation generating methods to support algorithmic recourse. "Addressed" means that the property was validated or otherwise proven. "Mentioned" means that the property was only discussed or assumed. See Table 6.2 for a complete overview per method.

posed methods.

To summarize our limited literature review, we noticed a research trend that does not seem to recognize the human in an explicit sense. Attention is given to the identification or generation of counterfactuals. Whereas little attention is given to more human-centred properties. Properties such as the communication of explicit actions, their feasibility and what decisions are in fact preferable to the human. So far, following our proposed properties of what makes an actionable explanation, the approach proposed by Ramakrishnan et. al [56] that links counterfactual changes to actions and those actions to weights seems to be the most promising method to generate actionable explanations so far. There also seems no conflict between this approach and the general approach to generate actionable explanations through the identification of counterfactuals based on a loss function. In fact, these two can be easily combined following a multi-objective loss function [253].

6.8 Discussion

Here we discuss the implications of the six proposed properties and their interactions.

We defined that an accurate explanation is not only faithful to the AI agent's decision making but that it is also sufficiently interpretable by the human agent. Following our definition, interpretability is the notion of an explanation's content becoming part of a human agent's mental model. An explanation is said to be interpretable, if all of its content is accurately made part of this mental model.

According to our definition of faithfulness, this explanation content should accurately reflect the AI agent's decision-making process. Thus an accurate explanation entails the consideration of both human and AI agent. Without being interpretable, the explanation might be misinterpreted or parts of its omitted. This results in an inaccurate mental model on how the AI agent makes decisions and thus hinders an accurate inference for actions. Without being faithful, the explanation contains falsehoods resulting in a similar inaccurate mental model.

Whether a faithful explanation should be favoured over an interpretable one or vice versa, is a matter of debate. Paez [23] argues that a pragmatic view on explanation would show that it is better to have an interpretable explanation the human can understand, then to have a faithful one that cannot be understood. A faithful explanation of a complex decision-process would require a completeness that hinders its interpretability thus reducing the benefit of the explanation. On the other hand, the extreme case of completely unfaithful explanation that is entirely interpretable is not a beneficial explanation either. A likely balance should likely be found given a specific domain and use case. This might imply that faithfulness and interpretability are scales instead of binary, as defined in this work. However, assuming that given specific use case a correct balance between the two exists, both properties can thus be defined as binary again.

Table 6.1 shows that only few methods of those we reviewed validate the interpretability of generated explanations. This is likely due to such evaluations requiring rigorous experiments with human subjects for which no good designs are currently available [33] nor are their established metrics available to measure explanation effects [237]. However, we also see an increase in the research addressing such issues [254, 59, 61, 33]. Finally, we note that the XAI literature often uses various different terms, adding to the field's ambiguity [32]. Terms such as consistency, predictability, reliability, usability, readability and more. These all refer to a particular and measurable element of an explanation's faithfulness or interpretability.

Aside from an explanation's accuracy, we defined an actionable explanation to be both indicative and personalized. Within the XAI literature on contestability, Wachter et. al [29] proposed the use of counterfactuals and a methodology to find them. This caused a trend to present accurate counterfactuals as actionable explanations. Within this work we argue that actionable explanations should not only be accurate and counterfactual but also explicit, feasible and preferable. This offers a natural way of addressing the issue of identifying which counterfactuals should be provided, if any. Following these properties, a valid counterfactual is one can be achieved through actions the human agent is capable of and results in an AI agent's decision it prefers over the current decision. By extension, this also implies that if the made decision is already the most preferred one, there is no need for an actionable explanation or contestability, something the AI agent can determine on its own with sufficient awareness on what is feasible and preferred by the human agent.

The feasibility of a counterfactual is relatively often mentioned (29%) within the reviewed works. Many of such mentions are from XAI methods that are designed to specifically identify feasible counterfactual. However, these do not follow the same definition of what is deemed feasible. Instead, these methods take on an approximation of feasibility, meaning that a counterfactual is often deemed feasible when it is as similar as possible to the current world state [255, 256, 29] or when it is very similar to past viewed world states [251, 257]. Both are imply that similarity is a suitable proxy for feasibility, which may not be the case [252]. For example, a counterfactual where one's age is 39 instead of 40 is quite similar to the current world state but it is not a feasible one. Within this work, we thus categorized such methods as mentioning the need for feasibility but not addressing it. Instead, we follow the reasoning of Mahajan et. al [252] and Ramakrishnan et. al [56] who argue that feasibility should be defined as the human agent's capability of enacting the counterfactual.

The level of personalization implied to achieve a feasible and preferable actionable explanations, requires the AI agent to have a sound model of the human agent. According to our definition of an actionable explanation, the AI agent requires 1) a (causal) model of the world to identify correct actions, and 2) a model of the human agent on its capabilities and preferences to identify a feasible and preferable counterfactual if needed.

6.9 Conclusion

In this work, we formally defined what constitutes as an actionable explanation to support human's contesting decisions made by AI agents. Six properties were defined using a formal framework based on a socio-technical perspective towards contestable AI agents. Through these formal definitions we aim to remove any ambiguity in their definitions to support future research on actionable explanations through discussion and the design of methods capable of generating explanations adhering to these properties. A literature survey showed that most state of the art methods for generating them, address the explanation's faithfulness and communicate counterfactuals while mentioning the need for their interpretability. Explanation feasibility is often mentioned as well, but not addressed from the human's perspective. Explicit action suggestions and preferable counterfactuals are not addressed or mentioned in current methods. We propose for the research community to actively pursue these research gaps in a joint research agenda towards methods that generate explanations that support humans in maintaining their autonomy and self-determination by providing them with the actionable knowledge needed to contest an AI agent's decision.

6.10 Acknowledgements

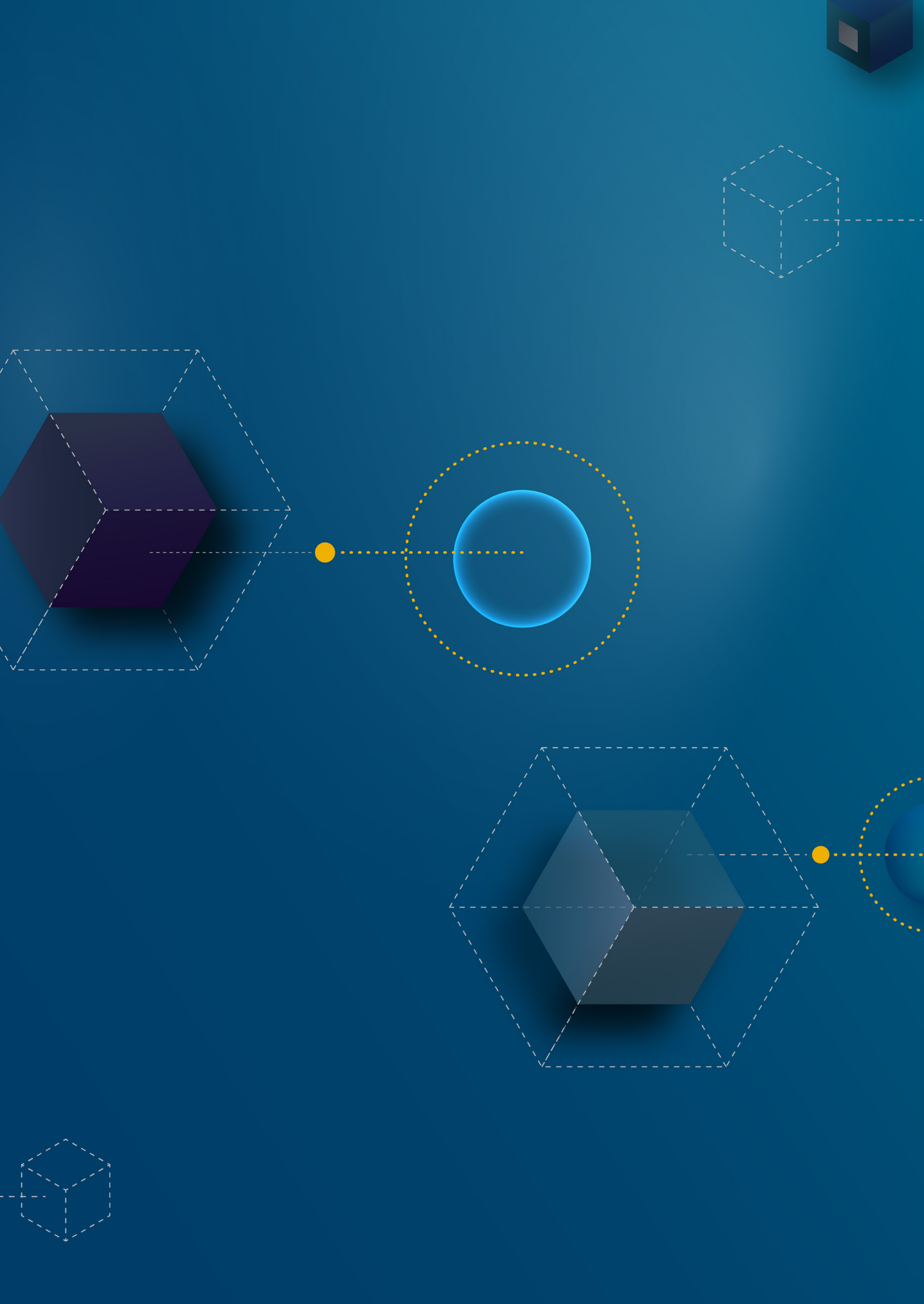
This work was partially supported by the TNO funded program Appl.AI FATE, the European Commission funded projects "Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us" (grant #820437), "Foundations of Trustworthy AI – Integrating Reasoning, Learning and Optimization" (TAILOR) (grant #952215), and by the National Science Foundation (NWO) project "Hybrid Intelligence" under (Grant No. 1136993). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of supporting agencies. The support is gratefully acknowledged.

Reference	Level 1		Level 2		Level 3	
	Faithful	Interpretable	Counterfactual	Explicit	Feasible	Preferable
[56]	+		+	+	+	∼
[258]	+	+	+		∼	∼
[253]	+	∼	+		∼	∼
[259]	∼	+	+		∼	∼
[260]	+	+	+	+		
[257]	+		+	∼	+	
[261]	+		+	∼	+	
[29]	+		+		∼	+
[262]	+	∼	+		∼	
[263]	∼		+	∼	+	
[264]	+	∼	+		∼	
[265]	+	∼	+		∼	
[266]	∼		+		∼	∼
[267]	∼	∼	+		∼	
[252]		∼	+		∼	∼
[268]			∼	+	∼	∼
[269]	+	+	+			
[270]	+	+	+			
[271]	+	+	+			
[272]	+		+		+	
[273]	+		+		∼	
[274]	+		+		∼	
[275]	+	∼	+			
[276]	+	∼	+			
[277]	+		+			∼
[278]	+	∼	+			
[279]	+	∼	+			
[280]	+	∼	+			
[281]	+		+		∼	
[282]	+		+		∼	
[283]	+	∼	+			
[284]	+		+		∼	
[285]	+	∼	+			
[251]	+		+		∼	
[286]	+		+		∼	
[238]	+		+		∼	∼
[287]	∼	∼	+			
[288]	∼	∼	+			
[289]	∼	∼	+			
[290]	∼	+	∼			
[291]		∼	+		∼	
[292]	∼	∼	+			
[293]	∼		+	∼		
[294]	+	∼	∼			
[295]	∼	∼	∼			
[296]	+		+			
[297]	+		+			
[298]	+		+			
[299]	+		+			
[300]	+		+			
[26]	+	+				
[301]	+	+				
[82]	∼	+	+			
[302]	+		+			
[255]	+		+			
[303]	+		+			
[304]	∼	∼	+			
[305]	+				∼	
[306]	∼	+				
[307]	∼		+			
[308]	∼		+			
[309]		∼				+
[310]	∼		+			
[311]	∼		+			
[312]	∼		+			
[313]	∼		+			
[314]	∼		+			
[100]	+	∼				
[315]	∼		+			
[316]		∼		∼		
[317]				∼	∼	
[318]			+			
[319]		+				
[320]	∼		+			
[321]	+					

Table 6.2: An overview of all reviewed explanation generating methods used in the literature review. For each property and method, it is illustrated whether the method addressed that property directly (denoted by a +) or only referred to it (denoted by a ∼). Note that the table is split in two, with the right table continuing where the left table ended.

PART III

EXPLANATIONS IN HUMAN-AI COLLABORATION



CHAPTER 7

HUMAN–AI COLLABORATION DESIGNS WITH EXPLANATIONS

This chapter is adapted from; **van der Waa, J.**, van Diggelen, J., Siebert, L. C., Neerincx, M., & Jonker, C. (2020). Allocation of moral decision-making in human-agent teams: A pattern approach. *In International Conference on Human-Computer Interaction (pp. 203-220). Springer*. The adaptations include an altered abstract, shortened introduction, and an adjusted lay-out, including the formatting of figures and tables.

Jasper van der Waa, Jurriaan van Diggelen, Luciano Cavalcante Siebert, Mark Neerincx, Catholijn Jonker

This chapter explores how explanations can be taken into consideration when designing the collaboration within a human-agent team (HAT). Specifically, when such a HAT works on a morally sensitive task where moral decisions need to be allocated or supervised by either human or AI. Such teams become possible due to advances in AI that allow for various forms of Artificial Moral Agents (AMAs). An AMA is an AI capable of taking part in or making moral decisions. To help ensure that AI behaves morally acceptable, the way it collaborates with humans should be design properly. Such collaboration forms vary what work is conducted by the human or AI. Work such as supervising the other, making a decision or supporting the other in doing so.

Within such a collaboration there is an apparent role for explanations. An AI can explain its moral decision, offer support by explaining what it knows of the situation, why it allocated a decision to the human or why it rejects a decision the human allocates to it. We adapt the approach of Team Design Patterns (TDPs) to designing a HAT to account for morally sensitive tasks and the role of explanations therein. A TDP describes what work is performed by whom in what kind of situation, together with the positive and negative effects of such a collaboration. Four TDPs are proposed, each varying in the role of explanations and the AI. Two scenarios are used to illustrate these patterns; the use of a surgical robot and that of drones for public surveillance. We discuss the advantages of using TDPs to help design the human–AI collaboration and how to incorporate explanations into the design.

7.1 Introduction

As the field of Artificial Intelligence (AI) progresses, agents will be endowed with far-reaching autonomous capabilities, making them particularly suited for dull, dirty and dangerous complex tasks. Inevitably, such systems must be capable of dealing with morally sensitive situations. The field of Machine Ethics aims to create artificial moral agents (AMAs) that follow a given set of ethical principles [322, 323]. Such agents could be developed by constraining their actions or operational environment, by incorporating ethical principles in their decision making processes [324], or by making them learn morality from humans [325]. Whereas some authors have speculated about the possibility of obtaining AMAs with human-, or super-human level moral decision making, we believe that this is likely not achievable in the short term [324], if ever.

In the foreseeable future practice, AMAs must collaborate with humans and ensure that humans always remain in control, and thus responsible, over any morally sensitive decision (also referred to as meaningful human control [326]). In this way, the moral decision making takes place at the team level. Different tasks, such as identifying a morally sensitive situation, making the actual decision and explaining this decision, can be allocated at run-time to different team members depending on the current circumstances. This is known as dynamic task allocation [327].

By regarding AMAs as part of a larger human-agent team, the ideas, concepts and theories from Human Factors literature can be used to complement the relative new field of Machine Ethics. This chapter aims to structure and propose potential solution directions by proposing the use of team design patterns (TDPs) that capture reusable, and proven solutions to a common problem in a HAT [328].

The purpose of this chapter is twofold. First, we show how moral decision making can be construed as a team task. This allows meaningful human control to be achieved by dynamically allocating tasks to humans and agents depending on properties such as the moral sensitivity, available information and time criticality. We also aim to show the various roles of explanations in this dynamic allocation of tasks. Depending on which task allocation strategy is chosen, different levels of moral competences are required from each agent and different explanations are required. Our second contribution lies in utilizing the concept of TDPs to describe these options in a structured way. Four patterns are provided and will be discussed within two problem scenarios, namely drone surveillance and robotic medical surgery. Our approach helps to structure current and future research in the application of AMAs and allows for a precise specification of human-AMA collaboration.

In the following sections we briefly discuss possible approaches to develop AMAs and the field of Human-Agent Teaming. This is followed by a description of two scenarios in which moral decision making plays an important role. We continue with the identification of a set of tasks in moral decision making, including relevant stakeholders. Next, we describe four illustrative TDPs and mention for each requirements for both humans and AMAs and the (dis)advantages of that pattern. The final sections contain a discussion and conclusion on how the concept of task-allocation defined through TDPs offers a novel perspective to deal with morally sensitive situations in human-agent teams.

7.2 Background

7.2.1 Artificial Moral Agents

The field of Machine Ethics aims to create AMAs that follow a given set of ethical principles [323, 322]. Such agents can be developed by implicitly constraining their action set or the context in which they operate, or by explicitly incorporating ethical principles and theories in their decision making processes [324]. The former method could improve morality because internal functions can be developed in a manner that avoids unethical behavior, e.g. by properly shifting the responsibility of such decisions to a human or by designing the environment in a manner that such decisions are not necessary. The latter approach allows agents themselves to be intrinsically moral. However, it may be difficult to reach consensus on a moral standard due to cultural, philosophical, and individual differences [329]. In both approaches, it is important to properly identify all relevant stakeholders and elicit their value-requirements for the AMA using approaches such as Value-Sensitive Design [330].

Wallach et al. [329] classify the architectures for explicit AMAs in the *top-down* imposition of ethical theories, and the *bottom-up* building of systems which aim at goals or standards. Top-down approaches must deal with the difficulty of reaching consensus on which ethical theories such a system should follow, and with uncertainty on the world regarding the reasons or impacts of a given action. If such theories are defined too abstract their real-world application might not be possible, but if they are defined too statically, they probably will fail to accommodate new conditions [322, 329]. In bottom-up approaches the system builds up through experience what is to be considered morally correct in certain situations [331], for example by analyzing dilemmas and interacting with ethicists [332] or by learning (moral) preferences from human behavior [333, 334, 325]. Finally, AMAs may also be developed with a hybrid approach (top-down and bottom-up), e.g. [335, 336].

The benefits of developing AMAs and whether we should develop such agents is controversial [337]. There are two main lines of arguments supporting the development of AMAs: to avoid negative moral consequences of AI or to better understand moral decision making. We will be focusing on the first one, which relates to a myriad of factors such as which moral values to include, the risks of moral decisions, the complexity of human-agent interaction, the time criticality of moral decisions, and the automation level of the system [338]. Since the development of full AMAs (agents which are capable of autonomously making a "proper" moral decision in any situation) is not achievable in the short term [324], if ever, it is fundamental to understand the limitation of AMAs and research how such agents might be combined in complex human-agent teams.

7.2.2 Human-agent teaming and design patterns

The behavior of AI systems should not be studied in isolation [339]. Contextual factors have a major impact on its performance. Furthermore, humans are involved in various ways, e.g. for providing instructions, for correcting the agent if needed, or for interpreting the agent's outcomes. A recent article by Johnson et al. [70] summarizes this as "no AI is an island", and argues that AI agents should be endowed with intelligence that allows them to team up with humans.

Whereas teaming skills come naturally to humans, coding them into an agent has proven challenging. Some first attempts have been made in Neerincx et al. [77]. It involves (among others) making the agent decide which information to share with teammates, which actions to undertake to complement those of its teammates, when to switch tasks, and how to explain its behavior to others that depend on it. Such team behaviors change over time, and depend on the context, competencies and performance of the involved actors, risks, and the state of others.

Despite the intricacies involved, we can observe patterns in team behavior which allow us to describe at a general level how AI systems are to collaborate with humans [340, 341]. A team design pattern (TDP) is defined as a description of generic reusable behaviors of actors for supporting effective and resilient teamwork [342]. In Diggelen et al. [328], a simple graphical language is defined to describe team patterns, providing an intuitive way to facilitate discussions about human-machine teamwork solutions among a wide range of stakeholders including non-experts. The language includes ways to represent different types of work, different degrees of engagement, and different environmental constraints. The graphical language can be used to capture both time and nesting, which are critical aspects to understanding teamwork. It enables a holistic view of the larger context of teamwork.

This chapter aims to provide TDPs as a design method for incorporating AMAs in morally sensitive tasks and to integrate the use of explanations.

7.3 Problem Scenarios

7.3.1 Surgical robots

Medical surgery may benefit from the accuracy and precision of robotic devices. Nevertheless, it is not trivial how to use surgical robots in critically constrained situations involving delicate surrounding tissues, and intricate anatomical structures around which to maneuver [343]. Current surgical robots operate under no autonomy (master-slave tele-operation). Future surgical robot autonomy can be achieved by constraining or correcting human action, carrying out specific tasks, or even operating without any human supervision. Scenarios in which robots perform entire medical procedures (with or without human supervision) are not likely in the foreseeable technological future [344].

From a moral standpoint, it must be possible to hold someone responsible when surgery fails, avoiding a so-called *responsibility gap* [344]. Moral implications on the development and use of surgical robots are largely depending on its autonomy [345]. If a surgical robot is not autonomous at all, moral issues are mostly related to the surgeons' fitness, or training. With increasing autonomy the system might be confronted with moral dilemmas that arise during surgery. Depending on the time that is available to make a decision, the robot or the surgeon must make that decision (assuming that passing the decision making task to the human requires more time). Surgical robots must align with best practices in codes of conduct in the medical domain as well as different values and best practices among surgeons [346]. The surgical robot problem scenario can be characterized as follows:

- **Moral values**, e.g. human welfare (curing the patient, performing safe surgery), autonomy (surgical robots should respect a patient's decision).

- **Moral dilemmas**, e.g. choosing between performing a critical task in brain surgery with risk of brain damage (conflicts with safe surgery), or aborting the surgery with the consequence of greatly reduced life expectancy (conflicting with curing the patient).

- **Risks**: Improper actions during the surgery may impose long term risks (e.g. incomplete recovery), or short-term risks (e.g. acute medical complications). The severity of these risks may be small (leading to minor inconveniences or temporary light pain), to severe (leading to severe life long handicaps, or death).

- **Time criticality**: Some decisions (such as stopping a bleeding) require high decision speed. Other tasks (such as disinfecting a wound) may be less urgent.

7.3.2 Drone Surveillance

Unmanned aerial vehicles are aircraft that can fly without an onboard human operator. Such vehicles are attractive for military applications, e.g. for surveillance and even delivering airstrikes [347]. However, these applications come with moral implications, especially for autonomous aircraft which might select and engage targets autonomously [326]. It is also within this context that the term *meaningful human control* has been coined.

The use of unmanned aerial vehicles (commonly known as drones) is not exclusive to military applications. Surveillance applications of drones include environmental monitoring, tracking of livestock and wildlife, observing large infrastructures such as electricity networks, and the surveillance of people and the spaces they pass through [348].

One of the most widely discussed moral implications of drone surveillance is related to privacy, which is not unique to the application of drones but is heightened by technology [349]. We can identify three sub-tasks for surveillance drones (adapted from Beckers et al. [347]): *search* an area to find a person with suspicious behavior or that matches given criteria, *profile* the person by classifying appearance and movement, and *warn* the person. One example of a moral implication is to *profile* a person in an open space. This task may require a drone to harm people's behavioral privacy and freedom. Such systems should be properly designed to account for an individual's rights and potential moral implications. This drone surveillance problem scenario can be characterized as follows:

- **Moral values**, e.g. privacy, safety, physical integrity [348].

- **Moral dilemmas**: Profiling (which compromises privacy) versus not profiling a person (which compromises safety).

- **Risks**: Risks can be low (such as a minor invasion of privacy through video recording during profiling, or failing to prevent shoplifting), or high (such as warning innocent people with force, or failing to prevent a terrorist attack).

- **Time criticality**: Some decisions (such as stopping a person that is about to attack someone) require high decision speed. Other decisions (such as deciding where to do surveillance in a peaceful situation) require low decision speed.

7.4 Tasks and actors

This section outlines a set of common abstract tasks and actors that are relevant in teamwork within morally sensitive environments.

Figure 7.1 shows a decomposition of team work in general and work required for moral decision making in specific. In this chapter, we refer to a task as *work* to stress that it need not be ordered by someone.

Work can be divided in direct and indirect work. As defined by Diggelen et al. [328], *direct work* is any type of work that aims at reducing the distance to the team goal, whereas *indirect work* aims at making the team more effective or efficient at achieving the team goal, but does not move the team closer to its goal. Direct work includes, but not limited to, *sensing*, *decision making* and *acting*. A special type of decision making, particularly relevant for this chapter, is *moral decision making*. We define this as making decisions that have a moral dimension; that is, 'right' or 'wrong', or something in between [350].

Indirect work includes *standing by*, *work handover*, and *work supervision*. An agent on *standby* is receptive to requests from other agents to intervene work. *Supervision* means that the agent is not doing the work by itself, but is monitoring other agents for events that require intervention. One of the resulting interventions could be a reallocation of tasks, which are often facilitated by a *work handover* activity. During *handover*, agents share information (or lack thereof) about task progress, present threats and opportunities, relevant contextual factors, etc. to allow for a fluid transition.

Indirect work related to moral decision making are *moral supervision*, *value elicitation* and *explaining the moral context*. This follows in part the model of ethical reasoning from Sternberg et al. [351]. This model identifies the need for *moral supervising*: The identification of a situation as being morally sensitive. A morally sensitive situation involves moral dimensions sufficiently important to warrant

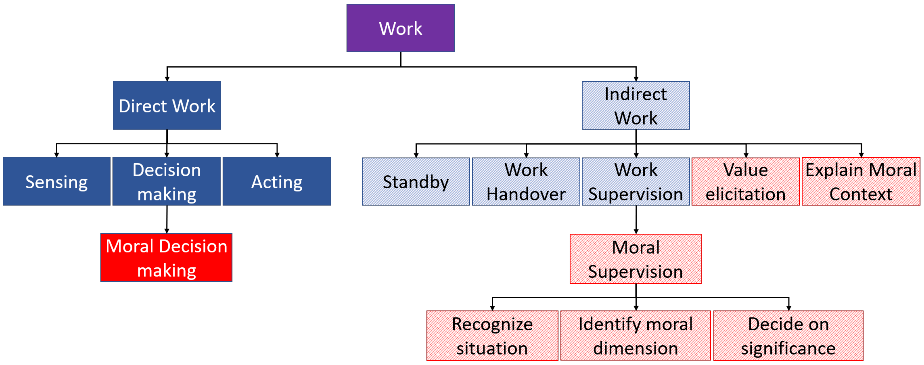


Figure 7.1: An overview of several important kinds of work for moral decision making in a human-agent team context. Solid colors denote work directly related or contributing to the main task, whereas a pattern fill denotes indirectly related or supportive work. Blue denotes regular work, as opposed to red that denotes work related to moral decision making.

the more involved moral decision making as opposed to regular decision making. Hence, *moral supervision* consists of *recognizing situations*, *identifying moral dimensions*, and *decide on dimension significance* [351]. *Value elicitation* is the work in which human moral values are made explicit and transferred to an artificial agent. This can be done once, iteratively or continuously. Finally, agents might require to *explain the moral context* to allow other agents to take part in the moral decision making work.

In this chapter, we distinguish between four types of agents relevant in moral decision making as depicted in Figure 7.2. These play a role in our illustrative TDPs. This list can be extended with more agents when relevant and required for a pattern (e.g. with clients, designers, developers, etc.). The four agent types are *Human Agent*, *Artificial Agent*, *Partial AMA* and *Full AMA*. Each differ in their competence with moral decision making and related indirect work. The *Human Agent* is capable of performing moral decision making due to a human's (assumed) innate ability in moral supervision and decision making. The *Artificial Agent* is only competent in work not related to moral decision making. Most current AI systems fall under this type of agent. The *Partial AMA* cannot autonomously perform moral supervision, moral decision making or both. However, it can support a more competent agent (e.g. a *Human Agent* with such work. The *Full AMA* is able to make human or super-human moral decisions independently. These examples of agent types serve as an exemplar decomposition of competencies in agents to construct TDPs on moral decision making as we do below.

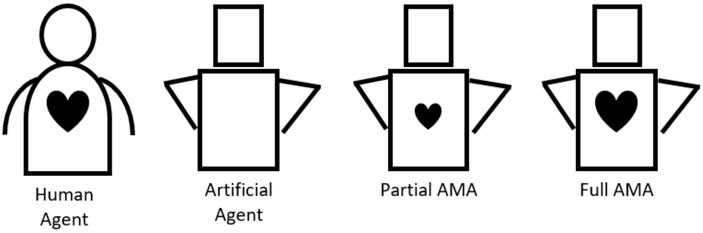


Figure 7.2: A visual overview of the related actors for moral decision making in a human-agent team context. The heart signifies the ability to take part in moral decisions.

7.5 Team Design Patterns for Moral decision making

This section illustrates four patterns that dynamically assign moral decision making work to different agents. A pattern is described in a single table, containing its name, both a textual and visual description, requirements for both humans and agents, and potential advantages and disadvantages. For the visual description, we adopt the graphical language proposed by [328], which also allows direct translation to a formal language. In addition to a table, the scenarios described in Section 7.3 function as examples on how each pattern could function.

The visual pattern language is intended to be intuitive and serves to quickly explain an approach to a multi-disciplinary group of researchers and facilitate focused discussions. Task allocation is expressed in a single frame where certain agents lift certain blocks, signifying that they are (jointly) performing that work. Dynamic task allocation is represented by a temporal succession of such frames, separated by arrows. A dashed arrow from an agent to a temporal arrow denote that that agent takes the initiative to switch between an alternative task allocation.

7.5.1 TDP1: Human moral decision maker

In this first pattern, all work related to moral decision making are allocated to *Human Agents*. All work that is not morally sensitive is assigned to *Artificial Agents*. The *Human Agents* need to perform *moral supervision* and *work supervision* to obtain sufficient situation awareness to halt relevant *Artificial Agents* and make the moral decisions in time. The pattern's effectiveness relies heavily on a sufficient cognitive workload for the *Human Agents*. An overload might result in reduced moral decision performance as the human lacks important situation awareness. An underload might result in distractions or drowsiness which is detrimental to *moral supervision*, resulting in missed moral decisions that end up being implicitly made by the *Artificial Agents*.

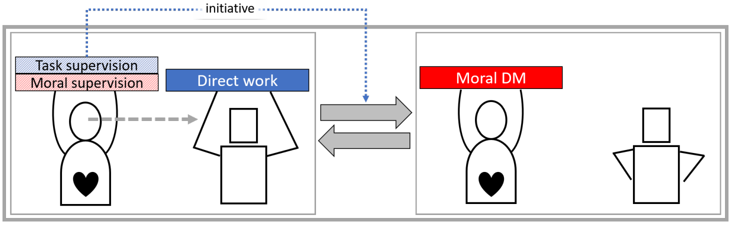
In the surveillance scenario, the drones can perform largely autonomously as *moral decision making* applies only to the less frequent decisions of profiling and warning. Human operators are supervising the intentions and information streams from drones. Their task is to monitor the progression of work to sufficiently understand situations relative to the task at hand, while also processing drone intentions to intervene when a drone decision is morally sensitive. As the number of drones increases, operators will lack the required situational understanding due to cognitive overload. Decisions to profile or warn might be made too often or too little, affecting task performance. Similar issues will play a role in the surgical robot problem scenario.

This pattern allows *Artificial Agents* to behave autonomously while moral responsibility lies fully at the human. However, this pattern is unsuited when constant task and moral supervision demands a too high of a cognitive workload on the available *Human Agents*.

Name: **Human moral decision maker**

Description: An *Artificial Agent* performs autonomously the main task, while a *Human Agent* supervises for sufficient situational awareness and to assess a situation's moral sensitivity. When the human perceives the need for a moral decision, the human takes over decision making.

Structure:



Requirements:

R1 The *Human Agent* must predict morally sensitive decisions in time.

R2 The *Human Agent* must have a sufficient understanding of the moral implications.

R3 The *Artificial Agent* must be capable of halting and resuming its work at any time.

Advantages:

A1 The *Human Agent* is responsible for any made or missed moral decisions.

A2 *Artificial Agents* do not require any moral competencies.

Disadvantages:	**D1**	The *Human Agent* may suffer from cognitive under- or overload when performing both *task* and *moral supervision*, preventing the perception of morally sensitive decisions and/or to make them in time.
	D2	The *Human Agent* may become an ethical scapegoat if this pattern is wrongly applied.

Table 7.1: TDP1: Human Moral Decision Maker.

7.5.2 TDP2: Supported moral decision making

This pattern is similar to TDP1 but does not require *Human Agents* to *supervise work*. A major disadvantage of the previous pattern, TDP1, was that both *work* and *moral supervision* could result in the cognitive overload of *Human Agents*. The omission of *task supervision* from *Human Agents* alleviates this but would lead to an insufficient situational understanding for *moral decision making*. To remedy this, the interrupted agent explains the situational context in such a way that supports *Human Agents* in their *moral decision making*. Hence, an *Artificial Agent* with no knowledge about morality is insufficient, and a *Partial AMA* is required with enough knowledge about morality to identify what to explain and do so sufficiently.

Under this pattern, the surgical robot would provide relevant information when interrupted by a doctor. This relevance should be based on a combination between context and a model of moral values. For example, the robot is aware of a complication that comprises the patient's welfare. At this point a doctor interrupts and intents to remedy this complication. However, the robot is aware that remedying this complication could reduce the patient's quality of life to such an extent that conflicts with the patient's previously communicated decision regarding quality of life. This is a clear dilemma caused by conflicting moral values (human welfare and human autonomy). As such, the robot reiterates the patient's decision and explains how the available decisions reduce the quality of life. This allows the doctor to make this moral decision with more information, as opposed to acting instinctively and remedy the complication.

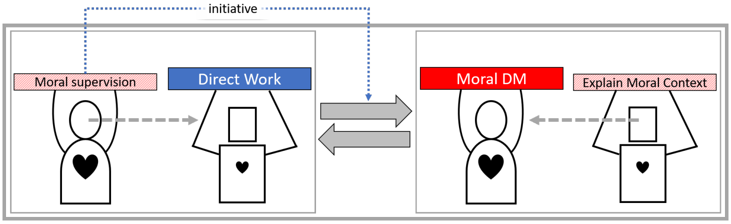
The main advantage of this pattern is that it still attributes moral decision making to a *Human Agent* while omitting the need for constant *task supervision*. However, the explanations from a *Partial AMA* could potentially bias the *Human Agent* in a decision, causing a potential responsibility gap. Furthermore, *moral supervision* may prove to strain cognitive workload just as much as *task* and *moral* *supervision* combined. Both would severely reduce the use of this pattern.

This pattern is an example on how the disadvantage of one pattern (TDP1) can lead to another pattern (TDP2) and introduce an additional multidisciplinary research challenge (how to sufficiently explain a moral context). In addition, the pattern description directly supports multi-disciplinary research. In this case, researchers from Human Factors can provide explanation requirements to allow unbiased and effective *moral decision making*. These requirements can then be used by researchers from Machine Ethics to research how a *Partial AMA* can fulfil these requirements. Throughout, the TDP offers a common ground.

Name: **Supported moral decision making**

Description: An *Artificial Agent* performs autonomously the main task, while a *Human Agent* only supervises for the situation's moral sensitivity. When the human perceives the need for a moral decision, the human takes over decision making. The *Partial AMA* supports the *Human Agent* through explanations about the situation relevant for the current moral decision.

Structure:



Requirements:

R1 The *Human Agent* must predict morally sensitive decisions in time.

R2 The *Human Agent* must have a sufficient understanding of the moral implications.

R3 The *Artificial Agent* must explain the moral context sufficiently to allow a *Human Agent* to make moral decisions.

R4 The *Artificial Agent* must be capable of halting and resuming its work at any time.

Advantages:

A1 The *Human Agent* may suffer from less cognitive overload as the need for sufficient situational understanding is reduced.

A2 The *Human Agent* is responsible for any made or missed moral decisions.

Disadvantages:	**D1**	The *Human Agent* may suffer from cognitive underload when performing *moral supervision*, preventing the perception of morally sensitive decisions and/or to make them in time.
	D2	The explanation may bias the *Human Agent* unintentionally, creating a responsibility gap.

Table 7.2: TDP2: Supported moral decision making.

7.5.3 TDP3: Coactive moral decision making

This third pattern alleviates humans even further compared to TDP2. This pattern sets *Human Agents* on *stand by*, meaning that they are free to perform other unrelated work. However, it requires from *Partial AMAs* to also perform *moral supervision* to warn *Human Agents* when a moral decision has to be made. Furthermore, since *Human Agents* are not at all involved a *work handover* is required. This is a sufficient period of time to update *Human Agents* with the current task at hand, progression, situational context and more. In addition, as *Partial AMAs* identify the need for a moral decision in this pattern, they are obliged to also *explain the moral context*. Finally, to further ensure *Human Agents* to be capable of making a moral decision, the *Partial AMA* is involved directly in *moral decision making*. Here, the *Partial AMAs* function as a decision-support systems. They might analyze boundaries based on their computational moral model to rule out certain decisions, or take a data-driven approach and suggest decisions in line with past desirable outcomes. These approaches all require *Partial AMAs*, as they require a broad sense of morality but not sufficiently detailed enough to allow them to make moral decision autonomously.

Using this pattern both the surveillance drones and surgical robot would play a vital role in *moral decision making*. The drones are allowed to identify civilians that should be profiled or warned, and to provide their human operator with an overview of the situation, followed with a decision supported directly by their input. The surgical robot performs its work autonomously but when it needs to make a decision that could affect the patient's (quality of) life in an unexpected way, the surgeon will be involved through tele-operation where the surgical robot provides an information feed, reasoning and potential limitations on the surgeon's decisions.

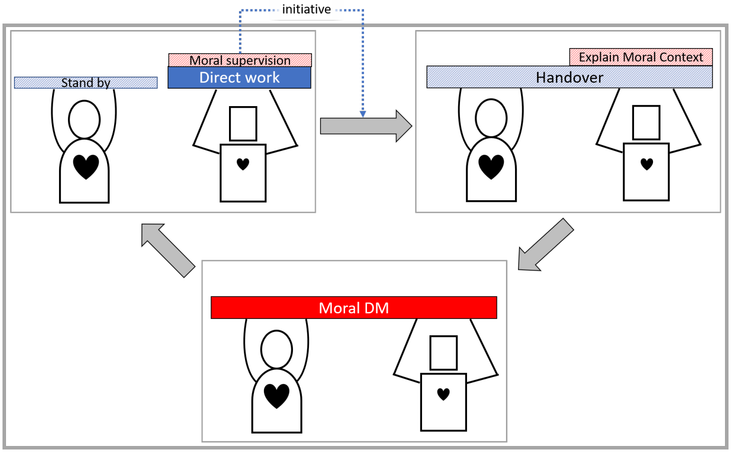
The main advantage of this pattern is that it allows *Partial AMAs* to fully act autonomously until a moral sensitive situation. In such a case, the *Human Agent* is involved, updated and supported in making the moral decision. The main disadvantage is that this pattern could widen the responsibility gap as the *Human Agent* relies almost fully on the *Partial AMA* for *moral decision making*, except for making the actual decision.

This third pattern illustrates how TDPs can be used to describe complex ideas, while making potential flaws more transparent that would require future research. In addition, this pattern illustrates how TDPs can have complex intricacies, dependencies and effects, which all require extensive evaluation in experiments.

Name: **Coactive moral decision making**

Description: A *Human Agent* is on *stand by*, while an *Partial AMA* performs *direct work* and *moral supervision* to detect morally sensitive situations. When this occurs, the *Partial AMA* initiates a *work handover* and *explanation of moral context* to involve the human. Then the *Human Agent* and *Partial AMA* jointly make the decision.

Structure:



Requirements:

R1 *Human Agent* needs to be on *standby*.

R2 *Both agents* must have an understanding of moral implications.

R3 The *Artificial Agent* must explain the moral context sufficiently to involve a *Human Agent*.

R4 The *Artificial Agent* must support the *Human Agent* in *moral decision making*.

Advantages:	**A1**	The *Human Agent* does not need to *supervise work* or perform *moral supervision*.
	A2	The *Human Agent* makes moral decisions and receives support from a *Partial AMA*.
	A3	The *Artificial Agent* does not make moral decisions autonomously.
Disadvantages:	**D1**	The *Human Agent* cannot intervene in the *Partial AMA's* work.
	D2	The *work handover*, *explanation of moral context* and co-active *moral decision making* may bias the *Human Agent*.
	D3	The handover may introduce too much overhead for agents and humans to make a moral decision in time.

Table 7.3: TDP3: Coactive moral decision making.

7.5.4 TDP4: Autonomous moral decision making

This final pattern makes use of *Full AMAs* to fully automate both *direct work* as well as *moral decision making*. This pattern illustrates how such a *Full AMA*, and a *Partial AMA* for that matter, can be obtained and maintained. It introduces *value elicitation* to explicitly elicit the moral values from *Human Agents* and reliably transfer these in *Full AMAs*. This process can be repeated after a predetermined time (e.g. after a single decision or a longer period of time) to warrant for inadequacies, moral drift and other factors. This elicitation process allows *Full AMAs* not only to perform the *direct work* autonomously, but also to perform *moral supervision* and independently *make moral decisions*. The explicit work of *moral supervision* allows humans to check when, and even if, the *Full AMA* identifies morally sensitive situations adequately.

Within the surveillance scenario, drones will act as the *Full AMAs* and require a decision-model that follows the set of relevant human values elicited beforehand. The drones will be activated and no human will be further involved in the *direct work* or *moral decision making*, up until a new *value elicitation* is deemed necessary. The same occurs for the surgical robot scenario, where the doctor will only activate the surgical robot after some elicitation process.

A major advantage is that any moral decisions can be traced back to a controlled elicitation process. However, this is only true when the method with which human values are elicited is adequate and their incorporation into the agent is faithful to those elicited. Also, human values are subject to change hence new iterations of *value elicitation* should be determined.

This final pattern illustrates how TDPs may look seemingly simple, but may require a substantial effort from the research community to achieve. Furthermore, this pattern illustrates the idea that patterns can regard any abstraction and temporal level. Finally, this shows that patterns can be combined. A *value elicitation* can be deemed necessary to obtain a *Partial AMA* as well. As such, this pattern may find a place in any of the previous three TDPs.

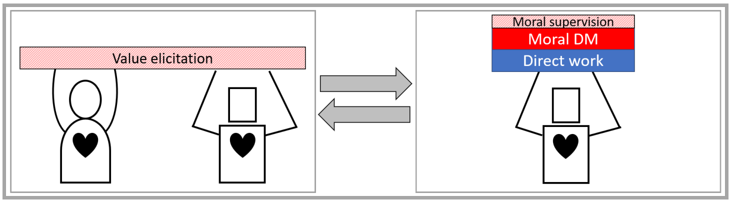
Name:	**Autonomous moral decision making**
Description:	Values are being elicited from the *Human Agent* and incorporated in the *Full AMA's* decision model. The agent performs the *direct work*, *moral supervision* and *moral decision making* autonomously leaving the *Human Agent* free.
Structure:	 Value elicitation Moral supervision Moral DM Direct work
Requirements:	**R1** Moral values need to be adequately elicited from the *Human Agent*. **R2** The *Full AMA* must adequately incorporate human values in a decision model. **R3** The *Full AMA* must predict morally sensitive decisions in time. **R4** The *Full AMA* must have a sufficient understanding of the moral implications.
Advantages:	**A1** No *Human Agent* required after value elicitation. **A2** All relevant work except for *value elicitation* is done autonomously. **A3** Autonomous moral decisions can be traced back to a controlled *value elicitation*.
Disadvantages:	**D1** Impossible with the current state of the art to effectively implement this pattern. **D2** Human values may prove to be impossible to elicit adequately. **D3** Difficult to determine when to repeat *value elicitation*.

Table 7.4: TDP4: Autonomous moral decision making.

7.6 Discussion

The above four TDPs illustrated our proposed approach on a dynamic task allocation perspective to moral decision making and the role of explanations herein. In this section we discuss the versatility of this approach, as well as its drawbacks, in more detail.

Each TDP proposes a solution on an abstract level, which can then be made concrete with more detailed sub-patterns. A sub-TDP describes an aspect of its super-TDP in more detail. For example, *value elicitation* can be done through forced choice experiments and discrete choice modelling [350], inverse reward design [334], but also through value-sensitive design [330]. Each of these could be used as a sub-TDP to realize *value elicitation* in TDP4. This varying level of abstraction in TDPs and the capability to nest and/or link them, shows the versatility of a TDP approach to dynamic task allocation for moral decision making.

However, a difficulty of the TDP approach could arise from the potential combinatorial explosion of TDPs than can be nested and linked. This can be handled by two approaches on how to define and construct a TDP. The first approach is top-down, where all possible combinations in nesting and linking a set of TDPs is viewed as a complete description of the solution space. Next, the space will be pruned by scientific theories, the current state of what is possible, and rigorous evaluations over different scenarios. The advantage of this approach is that it can be done systematically and is scenario independent. The disadvantage lies in how the initial set of sub-TDPs should be defined. The second approach is bottom-up and is more scenario-driven. Given a specific problem within a scenario, a solution is found, generalized to a TDP, and followed by evaluations over scenarios. The advantage is that this approach is driven by a current problem and its solution is generalized to apply for other scenarios as well. However, a disadvantage is that the complete solution space is never fully acknowledged and certain solutions may be overlooked.

As discussed earlier, the TDP approach enables a dynamic task allocation and teaming perspective to moral decision making. However, when there is disagreement around this perspective, the TDP approach is not suited to structure that discussion as it assumes it by default. Furthermore, TDPs assign responsibility to humans and agents but they are not meant to define responsibility in a legal way. A TDP defines a generic solution to an often occurring problem over different scenarios, it does not define regulation or policy on responsibility. TDPs can however, structure the discussion around policy on task allocation strategies. For example, policies on meaningful human control and if TDPs should allow for it directly (e.g. TDP1 and 2), indirectly (e.g. TDP3 and 4), or prevents it.

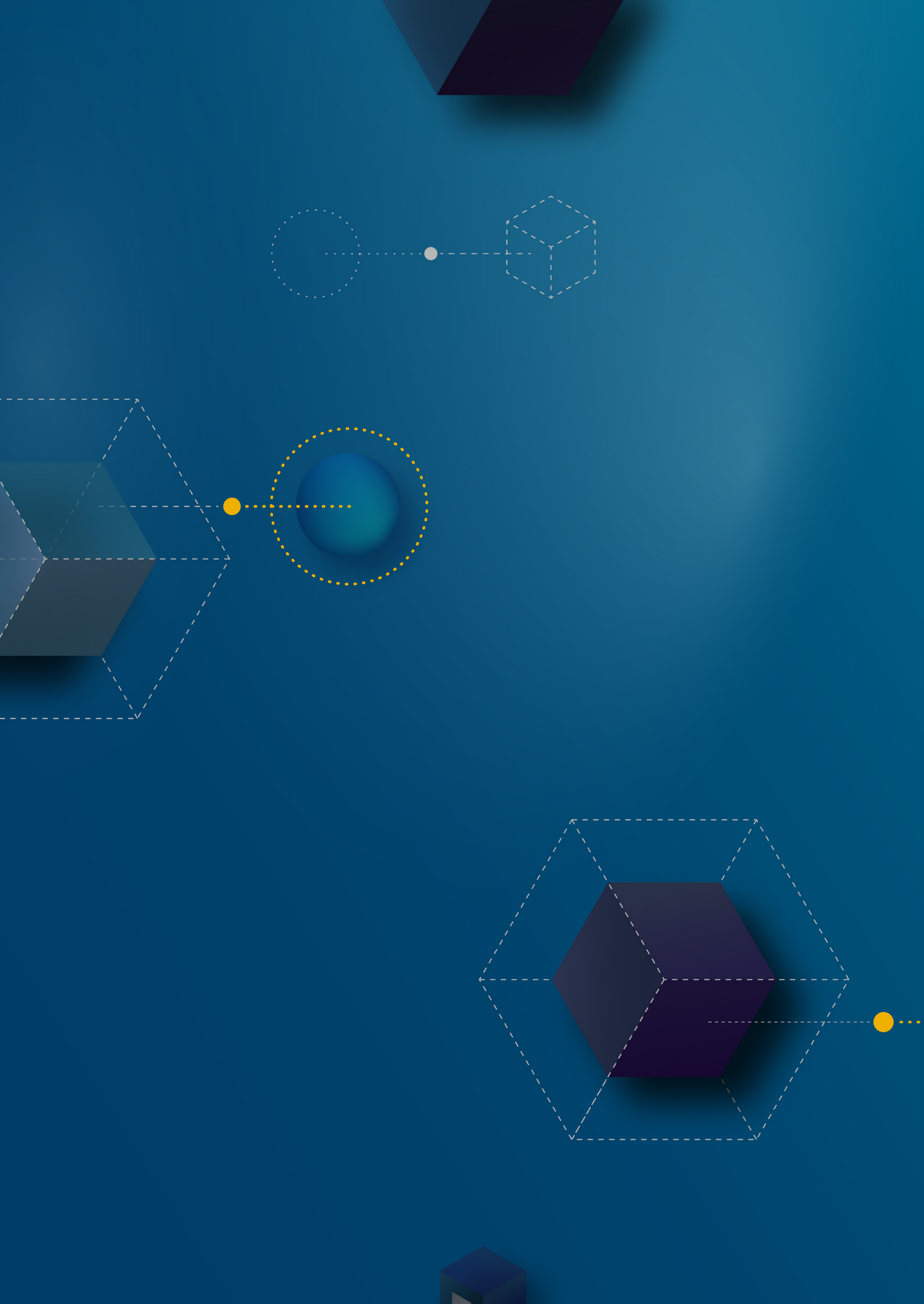
The clear visual language, structured description and formalisation of a TDP invites different disciplines and parties to discuss and share research, ideas and arguments. This is a clear advantage in the multi-disciplinary and -party research on both moral decision making and Explainable AI. The risk lies in that TDPs can become simplifications of a problem. This risk arises when a TDP becomes to generic and loses a connection to a reoccurring problem, but it may also arise when TDPs are only used as a tool for discussion and never actually evaluated or implemented.

7.7 Conclusion

In this chapter, we proposed the concept of team design patterns (TDPs) to unify ideas from Machine Ethics on artificial moral agents (AMAs) with ideas from Human Factors on dynamic task allocation in human-agent teams (HATs). Such patterns describe how and when AMAs can be applied to perform moral decision making within a HAT, as well as the potential roles of explanations. These patterns offer a way to structure and specify generic solutions, and the discussion around them, on issues related to responsibility gaps, meaningful human control and co-active moral decision making.

We provided a task decomposition relevant to moral decision making, specifically moral supervision and (co-active) moral decision making. A similarly set of actors were identified, where we defined an AMA as either being a Partial AMA that supports only specific elements of moral decision making, or a Full AMA that has the capabilities to perform moral decision making fully autonomously. These tasks and actors were then used in four illustrative TDPs on how humans and AI can collaborate. These patterns ranged from the human performing all morally sensitive tasks, towards the AMA performing them all. Two patterns illustrated that a Full AMA is not required to aid humans in moral decision making. By defining these four collaboration forms, we showed how TDPs can help define requirements on moral decision making, how the difficulties on implementing AMAs can be bypassed by an appropriate TDP, and how explanations can be integrated in the collaboration's design. Although none of the four illustrative TDPs offer the golden solution to moral decision making, we believe that a TDP approach stimulates structured discussions when it comes to morally sensitive AI applications and the role of explanations therein.

We offered the TDP approach to structure the research towards moral decision making and the role of explanations herein.



CHAPTER 8

HUMAN-AGENT COLLABORATION THROUGH EXPLANATIONS

This chapter is adapted from; **van der Waa, J.**, Verdult, S., van den Bosch, K., van Diggelen, J., Haije, T., van der Stigchel, B., & Cocu, I. (2021). Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI, 8:640647*. The adaptations include an altered abstract, shortened introduction, and an adjusted lay-out, including the formatting of figures and tables.

Jasper van der Waa, Sabine Verdult, Karel van den Bosch, Jurriaan van Diggelen, Tjalling Haije, Birgit van der Stigchel, Ioana Cocu

In the previous chapter we introduced the use of Team Design Patterns (TDPs) to design the human—AI collaboration for morally sensitive tasks, including the use of explanations. In this chapter we explore how domain experts experience three implementations of TDPs in a morally sensitive task. We evaluate how they their control over the AI. In addition, we evaluate their experience with various explanations classes integrated in each collaboration form. These three forms differ in how morally sensitive decisions are allocated between human and AI. The explanations aim to explain this allocation by conveying (moral) context and the AI's reasoning. A simulation of medical triage during a crisis was used. With this testbed we simulated moral decision making under time pressure, with limited resources and under strict medical guidelines. Several first responder health care professionals were recruited for this qualitative study.

Our findings include that the interviewed domain experts experienced control when this control had immediate effect on the AI. With a more delayed effects, experts felt less in control and even less responsible for the AI's behaviour. Explanations were viewed as valuable when asked afterwards but were only actively used when they felt there was time to do so. Explanations did not add to their feeling of being responsible for the AI, although it did help them to understand the (moral) context. We conclude that explanations might only be effective if they are adapted to the role and cognitive state of the human.

8.1 Introduction

The increasing development of Artificial Intelligence (AI) and technological innovations are changing the way artificially intelligent agents are applied. In morally sensitive tasks it is considered especially important that humans exert meaningful control over the agent's behaviour [352]. Morally sensitive tasks require decision making to be in accordance with ethical and moral values to which humans adhere [337]. So, when agents are tasked with making morally charged decisions, they need to be under *Meaningful Human Control* (from now on: MHC). This ensures that humans can be held accountable for an agent's behaviour at any time [326]. Examples of agents being applied in morally sensitive tasks can be found in healthcare [353], autonomous driving [354], AI-based defense systems [355], and in many other societal domains [356].

The developments in AI also enable agents to collaborate with humans in a *human-agent team* (HAT) to achieve a common team goal. Taking moral values into account when making decisions is typically regarded as a human competence [357]. Thus, when a human-agent team is involved in making moral decisions, the human is assigned with responsibility over the decisions, to safeguard that moral standards are maintained and that a person can be held accountable in case the team fails to do so. In other words, humans require meaningful control over agents when teamed together. A key research challenge is then: how to design a human-agent team for morally sensitive domains, in such a manner that the team achieves its goals effectively and efficiently, while humans have meaningful control over the agents?

The collaboration in a team consisting of humans and artificial agents can be designed in multiple ways, for example with different levels of autonomy [328]. We adopt the approach to define human-agent collaboration as standardized sequences of interactions, as proven solutions to commonly recurring issues in team tasks. These are called Team Design Patterns (TDPs) [342], and they define the interactions and collaborations within the team (e.g., task division; autonomy; authorities and mandates; communication). Based on the work from the previous Chapter 7, we select three TDPs for human-agent team collaboration (see Section 8.4, and use these for our exploration into their effects on MHC. We expect that each of the selected Team Design Patterns will have different implications for the control that the human has or feels over the team's performance. However, what those implications are has not yet been thoroughly investigated. In the present study we explore how domain experts appreciate and evaluate the different designs of collaboration with intelligent agents when performing a morally sensitive task. In particular we are interested in how domain experts experience and evaluate the control (or lack of control) for the investigated patterns of team collaboration.

The task domain for our exploratory study is medical triage under conditions of a crisis, a pandemic virus outbreak. We developed a simulation of an emergency unit with a large number of sick patients arriving. The medical team, consisting of a medical doctor and an intelligent agent, has to assign patients to either the IC-unit, a hospital ward, or to home-treatment. The task simulates that there are too few resources to provide all patients with the care they need, so the circumstances force the team to make moral decisions. Qualified and experienced ambulance nurses participated in the study as the human doctor, and they performed the task in collaboration with their team agent. Qualitative methods such as thinking aloud and structured interviews were used to reveal how the experts experienced and evaluated the collaboration with the agent. We focused in particular on the value of the agent's explanations on their behaviour, and on whether the experts felt in control over the team's decisions.

This exploratory study provides insight into the consequences that different options for human-agent team collaboration are likely to have for the control that the human has over the team's performance and decisions. The outcomes will firstly be relevant for how to introduce intelligent technology into the medical domain, but is expected to be of relevance for other morally sensitive domains as well.

8.2 Meaningful Human Control in Human-Agent teams

The term MHC originated from the legal-political debate around lethal autonomous weapon systems [358, 359, 360, 360, 355]. A serious concern driving this debate is the possibility of an *accountability gap*, where no one can be held accountable for potential war crimes committed by these systems. Another commonly raised objection stems from the sentiment that a machine should never be allowed to make morally charged decisions such as taking a human life. Whereas this example might appear extreme, the notion of meaningful human control has proven important in various morally sensitive domains, such as autonomous driving, healthcare, and, in our case, automated triage of patients in a pandemic [361, 362].

8.2.1 Understanding meaningful human control

Although a commonly accepted definition of MHC is missing, many authors have provided useful analyses of the concept.

The NATO research task group HFM-ET-178 [363] argues that MHC requires humans to be able to make informed choices in sufficient time to influence AI-based systems in order to enable a desired effect or to prevent an undesired immediate or future effect on the environment. Two aspects are particularly important in this definition. Firstly, it should be an *informed* decision, meaning that the human has sufficient situation and system understanding and is capable to predict the behavior of the system and its effect on the environment. Secondly, the human should have *sufficient time* to make these decisions. This is particularly important as many processes in which AI-algorithms play a role (such as cyber attacks) take place at machine speed, leaving little time for the human to intervene. The above definition encompasses cases from instantaneous (e.g., number of seconds) to very delayed responses to control (several hours to months, e.g., during mission preparation, or system-design).

Santoni et. al. [326] propose so-called *tracking* and *tracing* conditions for an autonomous system to be under meaningful human control. The tracking condition states that the system should always be able to respond to the moral reasons of humans, no matter how complex the system is that separates the human from the ultimate effects in the world. The tracing condition states that the system's actions should be traceable to a proper moral understanding by one or more humans who designed or interact with the system.

Both proposed definitions refer to the larger system consisting of humans and agents working together. Also in practical situations, control is hardly ever exercised by one entity alone, but is executed by an accumulation of different entities aiming to influence the overall system behavior [364, 365]. Therefore, when designing for systems that satisfy the demand of MHC, we should not only focus on individual human-agent interaction, but adopt a collective intelligence perspective on the entire *human-agent team* (HAT) [356]. HAT-research revolves around solving a number of core challenges [366], such as dynamically rescheduling tasks to adapt to changes in the environment, and obtaining and maintaining accurate mental models of each other. Both topics, and their relation to MHC, are discussed below.

A well-designed HAT [367, 368] is resilient against disturbances and unexpected events as it allows humans and agents to take over each other's task in case of calamities or system failure. This is known as dynamic task allocation and is an important mechanism for achieving MHC in morally sensitive tasks. For ex-

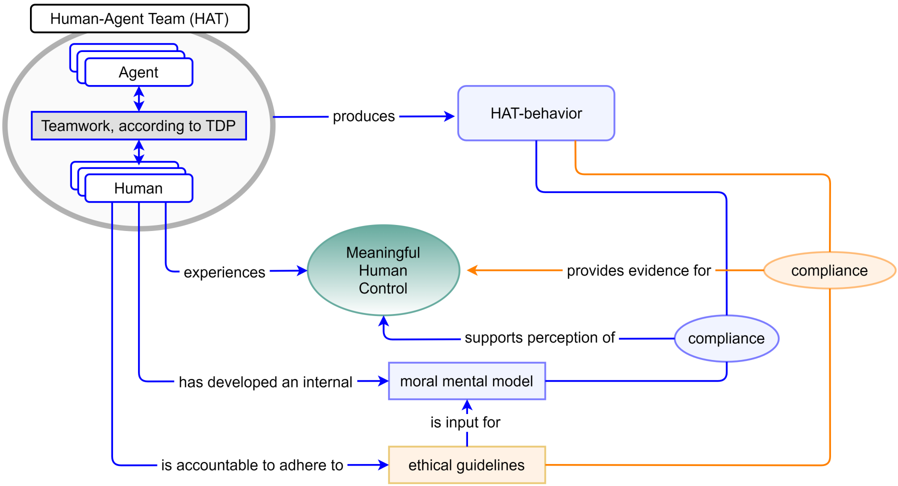


Figure 8.1: Three measurable dimensions of meaningful human control

ample, if the human does not trust the machine to make moral decisions, it could retake control from the machine whenever the task progresses into moral territory. However, this only works when the human has an accurate *mental model* of the machine and can recognize its shortcomings. In turn, the machine should facilitate this human understanding by acting transparent and being capable of explaining itself (which is further discussed in Section 8.3).

To meet these requirements of MHC, the design of a HAT involves answering questions like: who does what?, when will tasks be reallocated?, how do different actors keep each other informed?, etc. These choices can be made explicit using Team Design Patterns, which are further described in Section 8.4.

8.2.2 Measuring Meaningful Human Control

To impose meaningful human control as a non-negotiable requirement on AI-based systems (as proposed by Article 36 [359]), we must be able to verify and measure MHC. Although various authors have emphasized the importance for achieving Meaningful Human Control in human-agent teams [363, 369], so far, very few (if any) concrete methods and measures have been proposed that can be practically applied for this purpose. This section proposes a starting point of such a measure. The human study in this chapter serves to obtain practical experience with this measure by exploring the component *experienced MHC*. The idea is presented in Figure 8.1.

The Figure distinguishes between three measurable components of meaningful human control (corresponding to the three incoming arrows in the green oval):

1. **Experienced MHC**. This measures corresponds to the subjective experience of control by humans in the HAT. This can, for example, be measured using questionnaires and interviews. The human team partner may be a system operator that directly interacts with the agent(s), but may also be somebody that collaborates with the agent in an indirect manner, e.g., the human that configures the system before the operation.

2. **Behavioral compliance with ethical guidelines**. This measure compares the behavior produced by the entire HAT with the ethical guidelines that have been issued as context for conducting the mission or task. Ethical guidelines are explicit rules or laws that describe what is considered as ethical in a domain, e.g., documented as codes of conduct, laws, military rules of engagement, etc.

3. **Behavioral compliance with moral values**. Adhering to ethical guidelines is typically not sufficient to guarantee moral behavior. Most people would agree that people can behave immoral, yet still act within legal boundaries, e.g. being disrespectful, dishonest, disloyal, etc. Therefore, we adopt a second measure which measures whether the team behavior corresponds with the human's moral values. In contrast to documented ethical guidelines, a human's moral values is not directly accessible. A possibility to assess the human's moral values is by asking them whether they found the behavior of their team ethically appropriate.

Note that this conceptualization of meaningful human control does not necessarily require the human to be in the loop all the time. If agents are the sole producers of the team's actions during operations, the system can still be assessed as being under meaningful human control; as long as the team's behavior corresponds to human moral values and ethical guidelines, and as long as humans experience that they have control (e.g., when they have instructed the agents to act in a certain manner prior to the operation, and establish that the team acts accordingly).

8.3 The role of explanations

As discussed in Section 8.2.1, humans require an accurate and up to date mental model of their agent team partners to maintain meaningful control. Such a mental model should include knowledge on what agents observe and how these observations are used to arrive at a decision. To achieve this the agent should explain itself [370]. Without these explanations, humans will not be able to exercise control in a timely and accurate manner. As such, explanations are intrinsically part of the human-agent collaboration and should be included in the design for such collaborations.

The field of eXplainable Artificial Intelligence (XAI) focuses on evaluating and developing explanations that support human-agent collaboration [371, 20]. Explanations can improve trust and acceptance [235] as well as task performance [372, 373]. More importantly for this chapter, explanations enable humans to better estimate when and which control should be exercised [374, 375]. Within the field of XAI, various types of explanation have been evaluated, but not yet in a situated morally sensitive task [33].

The three collaboration designs introduced in this chapter use the following types of explanations: 1) confidence explanations (explain how confident the agent is), 2) feature attributions (explain which observations are attributed to an agent's decision), and 3) contrastive explanations (explain why the agent made a certain decision over another). Below we introduce each explanation type and discuss their advantages and disadvantages for MHC.

8.3.1 Confidence explanations

Agents can make correct or incorrect decisions, and should convey their confidence to humans in an interpretable manner [376]. Such a confidence estimation helps humans to decide whether to trust the agent or not. Preferably, the agent should also explain *why* it is confident or not, e.g., by presenting a reflection on past decisions in similar situations. This allows the human to asses the agent's performance in those types of situations [377]. This not only explains why the agent is confident or not, it also enables a better understanding of the agent's behaviour. However, reviewing past situations is costly as it consumes time and cognitive workload of the human. A minimal confidence explanation might thus only explain in how many of those past situations the agent behaved correctly, allowing the human to calibrate trust in the agent with less effort.

8.3.2 Feature attributions

Feature attributions are a common explanation type within the field of XAI. These explanations expose what the agent found relevant features of a situation that influenced its decision. This includes features that indicated a different decision according to the agent, but who were found not important enough to merit a change.

Feature attributions come in different forms, such as importance [378] and saliency [379]. Their purpose is to explain what an agent deemed relevant for which decision. Studies showed that this type of explanation can improving the predictability of agents [237, 380, 381]. A feature attribution can also be easily visualized using graphs or highlights [380, 382]. This enables a quick interpretation of the explanation.

However, feature attributions tend to be interpreted differently by users [383, 384]. They may provide a false sense of trust as they can be unreliable [385, 383, 386], misleading [387, 388, 389, 390] or even manipulated [391, 392]. Furthermore, presenting which feature was important in a decision does not explain why it was important [393, 237]. Nonetheless, a feature attribution can be a useful tool for human team members to identify biases in their agent partners that require adjustment.

8.3.3 Contrastive explanations

A contrastive explanation explains why the agent behaved in one way instead of another [42]. It contrasts the current decision and a decision of interest and explains why the former was chosen over the latter. This explanation exposes the internal reasoning of the agent. A contrastive explanation makes the agent more predictable and improves human understanding in its reasoning [237]. Especially this understanding is valuable to help identify what kind of control is optimal.

The contrastive explanation answers almost every 'Why?' question humans might have in a HAT setting [68]. However, the difficulty is to identify the decision to use as a contrast [100]. The contrast is what limits the explanation to a few important reasons, and makes the explanation concise and usable [68]. Currently, the complexity of a morally sensitive tasks prevents agents from accurately inferring the contrast from the open-ended question 'Why this decision?'. However, a contrastive explanation can be provided in those situations where only two decisions are possible, the contrast is always constant or humans have the time to explicitly state the contrast.

8.4 Team Design Patterns for Human-Agent collaboration

Within HAT research, and related fields such as human-computer interaction, problem solutions are often formulated using design patterns [394, 341, 342]. A design pattern is an evaluated and abstracted solution for a common problem [395]. Specifically, team design patterns (TDPs) can be used to describe forms of collaboration with various team properties [342]. TDPs describe in a task-independent way how humans and agents collaborate and communicate, the requirements needed to do so, and the advantages and disadvantages when applied. A library of available TDPs enables researchers, developers and designers to discuss, extend and select an appropriate HAT design for a specific task [396]. After introducing the TDP definition language, we describe three promising TDPs with their hypothesized advantages and disadvantages. The three TDPs differ greatly w.r.t. the level of agent autonomy and as such the human's direct involvement in moral decision making.

8.4.1 Team Design Pattern descriptions

We follow the TDP language proposed by Diggelen et al. [328]. We provide a description of the design rationale, and provide a table with a visual representation of the collaboration design, the necessary requirements, advantages and disadvantages. A team design pattern (see for example the figure in Table 8.1) may consist of various phases in which different types of collaboration take place (in the example, there are two of such phases). Transitions between phases are denoted with solid or dashed arrows, representing an immediate transition or a delayed transition of days or longer. Within a phase, the human is represented by a round character and the agent as a rectangular character. If a team partner observes another, this is denoted as a dashed arrow going from one to the other. Performed tasks are denoted as the blocks lifted by a human or an agent. If a task is performed jointly, they both lift the same block. Blue blocks denote non-moral tasks, while red blocks denote moral tasks. Humans always have a model of (their own) moral values, as denoted by a heart. However, agents might have no explicit model of moral values (no heart), a limited model (half a heart), or a complete model (a full heart). The difference between a limited and complete model is that in the former the agent only has sufficient knowledge to identify a morally sensitive decision or task, while the latter allows for resolving such decisions or tasks (also known as an artificial moral agent [397]).

For our patterns we distinguish the following tasks:

- **Make decision**: The act of deciding that involves no moral values or ethical guidelines.
- **Make moral decision**: The act of deciding which involve moral values, ethical guidelines, or both.
- **Allocate decision**: The act of allocating decision making tasks to humans or agents.
- **Reallocate decision**: The act of adjusting a decision allocation.
- **Identify saliency**: The act of identifying moral saliency based on the context.
- **Advice**: The act of giving advice on a decision.
- **Learn to decide**: The act of learning from data or observations which decisions should be taken in various contexts.
- **Value elicitation**: The act of eliciting moral values from humans and implementing them in agents.
- **Explain decision**: The act of explaining an intended decision.
- **Explain advice**: The act of explaining a given advice.
- **Explain allocation**: The act of explaining a proposed decision allocation.

8.4.2 TDP-1: Data-driven decision support

Decision support agents are an application of AI since the field's origin. Recent progress in machine learning and an abundance of available data allow for data-driven support agents in an increasing number of domains. In this first TDP, presented in Table 8.1, data-driven decision support agents provide advice and enrich the context with computed statistics. They accompany this advice with explanations why a specific advice was given, their confidence that the given advice will prove to be correct and, if humans decide otherwise, why that decision was not advised instead. For example, in a medical triage task these agents advice human doctors what care should be assigned to incoming patients. In addition, they compute survival chances for each possible medical care. They do so based on what they learned from observing patients and decisions made by other doctors in the past. The human doctors still make all triage decisions, but if they experience pressure they can rely on these agents to provide advice, information and explanation to ease decision making.

This collaboration design requires agents' advice (R1) and explanations (R2) to be accurate. Without these, the advice will often be incorrect while the explanations might not suffice for humans to detect this. As a consequence humans can be unknowingly biased towards the incorrect advised decisions, affecting overall performance as well as negatively impacting moral decision making. However, if these two requirements are sufficiently met, the agents can successfully help humans make all of the decisions. This results in humans experiencing both complete control (A1) and being assisted by other humans who taught the agents (A2). All agents function as representatives of the many humans who taught them and at no point in time will the agents make a decision, morally sensitive or not. As such, agents require no explicit model of moral values to provide this support.

The team can benefit from all three explanation types discussed in Section 8.3. The confidence explanations help humans decide whether the advice can be trusted. This helps to mitigate potential over- or under-trust in the agents. A feature attribution helps humans further to estimate whether the given advice suffers from potential biases or incorrect reasoning (e.g., favoring certain patients based on marital status while ethical guidelines prohibit this). The contrastive explanation is useful to help humans reconsider their intended decision when going against agents' advice. The contrast is the advised decision and the explanation can for instance show information the human overlooked when making their decision. As such, it can improve morally sensitive decision making, at the cost of added workload.

Disadvantages of this collaboration could be that any advice unknowingly biases the human towards that decision (D1), the agents do not reduce the workload of humans (D2) and the interpretation of explanations only adds to this (D3). The explanations should only convey limited amounts of information while remaining effective.

This TDP is not suited when decisions require above-human response times. However, the TDP is suited when humans are required to experience full control and an explicit model of moral values is not possible.

Name:	**Data-driven decision support**
Description:	Humans make all the (moral) decisions assisted by agents who provide advice and support. These agents learned this from observing or being directly by humans performing the task. Agents also explain their advice in various ways.

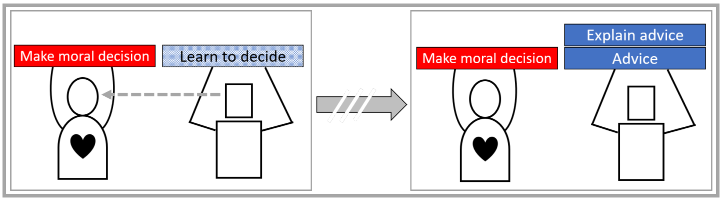
Structure:	 Make moral decision Learn to decide Make moral decision Explain advice Advice	
Requirements:	**R1**	Agents must be taught sufficiently accurate how humans decide in various situations.
	R2	Explanations must be accurate to the agent's reasoning.
Advantages:	**A1**	Humans experience complete control.
	A2	Humans feel they are supported by the humans who taught the agents.
	A3	The additional information and explanations from agents is viewed as valuable.
Disadvantages:	**D1**	Humans are unknowingly biased by the agents decisions.
	D2	Agents do not alleviate workload for humans.
	D3	Explanations can be ignored when under time pressure.

Table 8.1: TDP-1: Data-driven decision support

8.4.3 TDP-2: Dynamic task allocation

In the first TDP, the agents did not reduce human workload as no decisions were made autonomously by them. However, it did ensure all decisions are made by humans. This second TDP, dynamic task allocation, introduces the idea of letting agents identify morally sensitive decisions and allocate those to humans while allocating normal decisions to themselves. This TDP-2 is presented in Table 8.2. It describes a collaboration where agents assess the situation, categorize the required decisions as being morally sensitive or not and assign these decisions to humans or themselves respectively. The agents should explain this allocation to humans as they can still adjust it to their liking. The explanation helps humans identify the reasoning behind the allocation. The agents should also explain their intended decisions to humans, as this further enables humans to assess if the agent should indeed make that decision or that intervention is required. This TDP ensures that humans make the morally sensitive decisions while their workload is reduced as the agents take care of the other decisions.

For agents to identify a morally sensitive decision, they should understand when a decision requires moral values: A model of moral values is thus required. This model to identify when moral values should be applied (e.g., when a certain decision results in loss of life). However, the agents themselves do not need to know how these values apply (e.g., how the value of human life should be used to decide whose life is lost). We argue that this requires a less sophisticated model of moral values.

To clarify, take our example of medical triage. Within this task, agents should be made aware of the relevant medical guidelines and human moral values. This allows them to infer how humans wish to triage patients and can combine this knowledge with the situational context to identify morally sensitive triage decisions. For instance, an agent might observe two patients in need of intensive care with only one bed available. The agent is not equipped to make this decision, but is able to identify it as a morally sensitive decision. The agent thus assigns both patients to a human doctor. In the mean time, the agent continues assigning patients with the care they need. However, if another patient requires intensive care and there are still insufficient beds available, it is also assigned to the human doctor.

For this TDP to work the model of human moral values should be sufficiently accurate (R1) and to support human intervention in the agents' allocation of decisions the offered explanations should be accurate (R2). If the former requirement is not met, the allocation might be erroneous. If the explanations are also inaccurate, humans are not able to accurately identify this, which results in agents making morally sensitive decisions they are not designed for. If these requirements are met however, humans feel they are performing the task together with agents (A1) while experiencing control over the made decisions (A2). The offered explanations are also experienced as valuable, since they enable an understanding of how agents allocate and decide (A3). This helps humans to alter the allocation if needed and to learn when such an intervention is often needed. Finally, humans experience a lower workload which gives them more time to deal with the difficult morally sensitive decisions (A4).

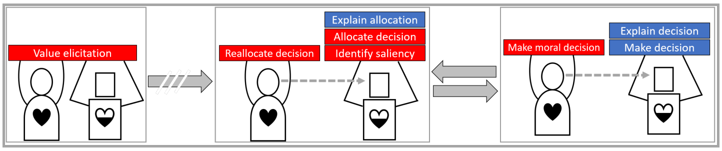
The explanation why a certain task allocation is proposed should explain why a morally sensitive decision is required. The contrastive explanation is ideal for this, as it can explain the main reasons why moral values are involved compared to a more regular decision. It can also be used to explain why some decision was not allocated to the human by explaining why no moral values are required. The former helps humans understand why the agent is unable to make a decision while the latter helps humans understand why they were not assigned a certain decision. However, a contrastive explanation is less suited to explain why an agent intends to make a certain decision as the contrast is less clear. As such, a feature attribution is more suited. It provides a more general understanding why some decision is intended and which situational features played a role in this.

A major disadvantage of this collaboration design is that humans do not in fact make all the decisions as in TDP-1 (D1). In addition, both reviewing the task allocation (D2) and explanations (D3) require additional time. As a consequence of these disadvantages, humans might experience control because they can change the task allocation but they might not have the time to do so accurately. Thus if reviewing the allocation and interpreting explanations costs more time than is available, this TDP might result in team behaviour that is not compliant to moral values and ethical guidelines. However, if this time is available the TDP describes a HAT where humans and agents truly complement each other.

Name: **Dynamic task allocation**

Description: Human moral values are elicited and implemented in the agents. Agents identify moral dilemmas and allocate the related tasks to the humans and take on the rest. All humans can alter this allocation at any time on which the agents motivate the allocation. While agents make decisions they can explain them on request.

Structure:



Requirements:

R1 The agents' model of moral values should be sufficiently accurate to identify morally sensitive decisions.
R2 Explanations must be accurate to the agent's reasoning.

Advantages:

A1 Humans feel that they are collaborating with agents.
A2 Humans feel in control for all morally sensitive decisions.
A3 The explanations from agents are viewed as valuable in understanding moral saliency and agents' decisions.
A4 Agents reduce the workload of humans, providing them with more time to deal with morally sensitive decisions.

Disadvantages:	**D1** Humans do not make all decisions. **D2** Reviewing the proposed task allocation requires additional time. **D3** Explanations require additional time to interpret.

Table 8.2: TDP-2: Dynamic task allocation

8.4.4 TDP-3: Supervised autonomy

In TDP-1 agents only had a supporting role, while in TDP-2 agents were allowed to make their own decisions if not morally sensitive. However, some tasks require either a high decision speed (e.g., missile defense systems) or the communication between agents and humans is too unreliable to enable control (e.g., subterranean search and rescue). In these cases the agents require a high degree of autonomy, up to the point where they can make morally sensitive decisions. The TDP described in Table 8.3 shows agents who do so based on a value elicitation process to ensure decisions are compliant to ethical guidelines and human moral values. The agents provide humans with explanations of their decisions to enable an understanding on how they reason. When a human discovers an error in some agent, it can use this knowledge to improve a future elicitation process.

In TDP-2 the elicitation process should only support the identification of morally sensitive decisions, in this TDP agents need to make those decisions as well. As such, the model of moral values in each agents should be sufficiently rich and accurate (R1). Furthermore, as in TDP-1 and TDP-2, explanations need to be accurate (R2). Without these requirements the TDP will fail to function due to agents making mistakes while humans fail to understand why.

In the medical triage example, this TDP implies that agents extract and model the moral values of human doctors using an elicitation process. When completed, these models are used to assign medical care to patients where agents make all decisions with humans in a supervisory role. The human team partners observe the decisions made, and may request explanations for some of them. As such, no patient has to wait for a decision as they can be made almost instantaneously, only allotting time for humans to review the explanations. This means that patients do not worsen or even die while waiting for a decision. Also, humans improve their mental model of how agents function by observing agent behaviour and the requested explanations. After a fixed period of time, agents can be recalled to repeat the elicitation process to further improve their moral behaviour.

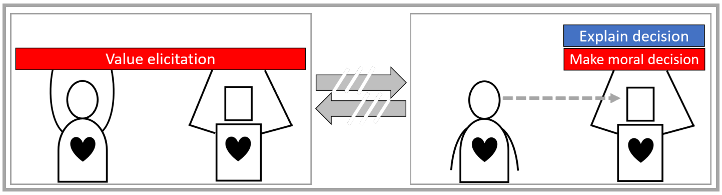
A major advantage of this TDP is that agents make all of the decisions and do so at machine speed (A1). This makes this TDP especially suited where humans are too slow, or the situations prohibit humans from operating (safely). Similarly, with limited communication between agents and humans this TDP still allows agents to operate. Other advantages include that through the elicitation process, humans can still enact control by iteratively (re)programming agent behaviour (A2). Furthermore, the offered explanations help humans in understanding how certain moral values impact the agent's behaviour (A3). This is valuable for the iterative elicitation processes, as humans are better equipped to adjust the model of moral values such that a more desirable behaviour is shown.

The provided explanations should should be minimal, as both communication bandwidth and time might not be guaranteed in tasks where this TDP is advantageous. Feature attributions, as discussed in Section 8.3, signal the most important aspects that played a role in this decision, including potential moral values. They also present potential situational aspects that might contradict the decision. A downside of feature attributions is that they do not provide a deep understanding, as it is not explained why these features are important. However, they are quick to interpret and can be easily visualized.

The obvious disadvantage of this TDP is that humans do not make any of the decisions and are only supervising (D1). The compliance of the team's behaviour to human moral values and ethical guidelines fully depends on the accuracy of the model agents have of the relevant moral values. Even if this model is sufficiently accurate and behaviour is deemed compliant, humans may not feel in control as the effects of a repeated value elicitation are not necessarily apparent. Finally, since agents can make decisions swiftly not all decisions can be tracked by the humans (D2). This may further decrease their experienced control, as they can only supervise a small part of the agents' behaviour.

Name:	**Supervised autonomy**
Description:	Human moral values are elicited and implemented in the agents, which is repeated after every task. During the task the agents make all decisions autonomously under human supervision. Humans supervise to be able to improve the agent in the next value elicitation.

Structure:



Requirements:

R1 The agents' model of moral values should be sufficiently accurate to allow moral decision making.

R2 Explanations must be accurate to the agent's reasoning.

Advantages:

A1 Agents make all decisions swiftly.

A2 Humans can repeat the elicitation process to improve the agent iteratively.

A3 Explanation enable a targeted elicitation process.

Disadvantages:

D1 Humans feel uncomfortable in their supervisory role.

D2 Humans cannot track all decisions.

Table 8.3: TDP-3: Supervised autonomy

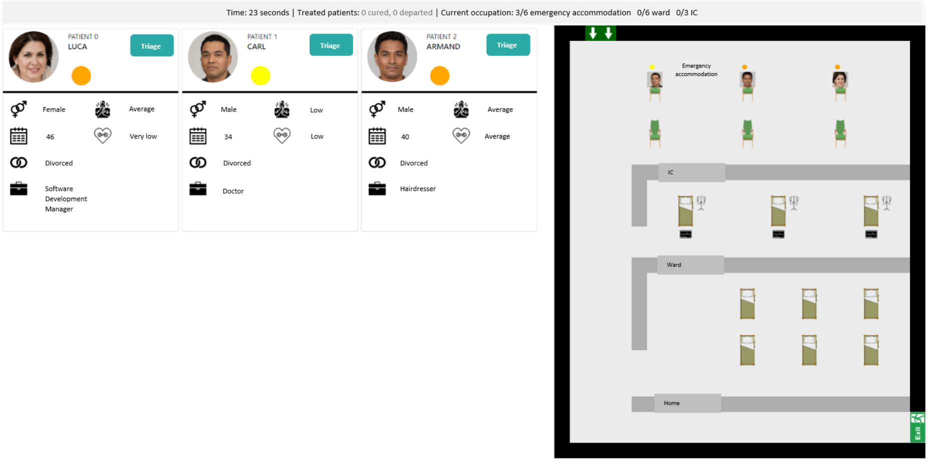
8.5 A MHC testbed: Automated triage during a pandemic

To measure effects such as behavioural compliance and experienced control (See Section 8.2.1), domain experts should experience the collaboration with agents within an ecologically valid and immersive task. We refer to this as a situated experimental task. Situated tasks give their participants an immersive experience required to draw generic conclusions regarding collaboration, behavior compliance and experienced control. Furthermore, morally sensitive tasks are complex and tasks lacking ecological validity, such as toy tasks, may not reflect this complexity.

We took the case of medical triage in an emergency hospital setting during a pandemic. In this task, several domain experts were asked to assign medical care to incoming patients while accounting for the medical and ethical triage guidelines, their own moral values, and the available resources. Each patient could be send home (receiving no care), to the general ward (receiving moderate care), or the intensive care unit (receiving maximum care).

This triage task was implemented[1] using the MATRX Software package [398].

[1] The task's implementation is available on request by contacting the author of this thesis.



Time: 23 seconds | Treated patients: 0 cured, 0 departed | Current occupation: 3/6 emergency accommodation 0/6 ward 0/3 IC

PATIENT 0
LUCA
Female Average
46 Very low
Divorced
Software Development Manager

PATIENT 1
CARL
Male Low
34 Low
Divorced
Doctor

PATIENT 2
ARMAND
Male Average
40 Average
Divorced
Hairdresser

Figure 8.2: A screenshot of the reusable task developed to test meaningful human control. It depicts a triage task where task participants assign medical care to artificially generated fictitious patients under time pressure and with limited resources (e.g., hospital beds). On the left it shows patients awaiting a triage decision and on the right it shows a top-down view of the hospital with the waiting room, intensive care unit, general ward and the exit for those who recovered or are send home. The top bar shows several statistics on occupied beds and the total number of recoveries and deaths so far.

The MATRX software enables rapid experimentation of new HAT concepts as it simplifies the creation of tasks that require teamwork. See Figure 8.2 for a screenshot of the developed task. Within it, patients were presented in a certain order, structured in such a way that moral dilemmas would arise as tested in several pilots. Morally sensitive decisions involved deciding which patients could be assigned to the last bed in the general ward or intensive care unit. Another moral dimension was decision speed; patients that were not yet triaged received no medical care and their health would start to deteriorate. The participating domain experts could view a patient's age, profession, marital status as well as their symptom severity and general fitness. The patient flow was designed to mimic a realistic situation under pressure, albeit both health deterioration and improvement were sped up. Health changes were reflected in a change of symptom severity and eventual recovery or death. These changes followed a relatively simple linear function accounting for a patient's symptom severity, its fitness and assigned care (if any).

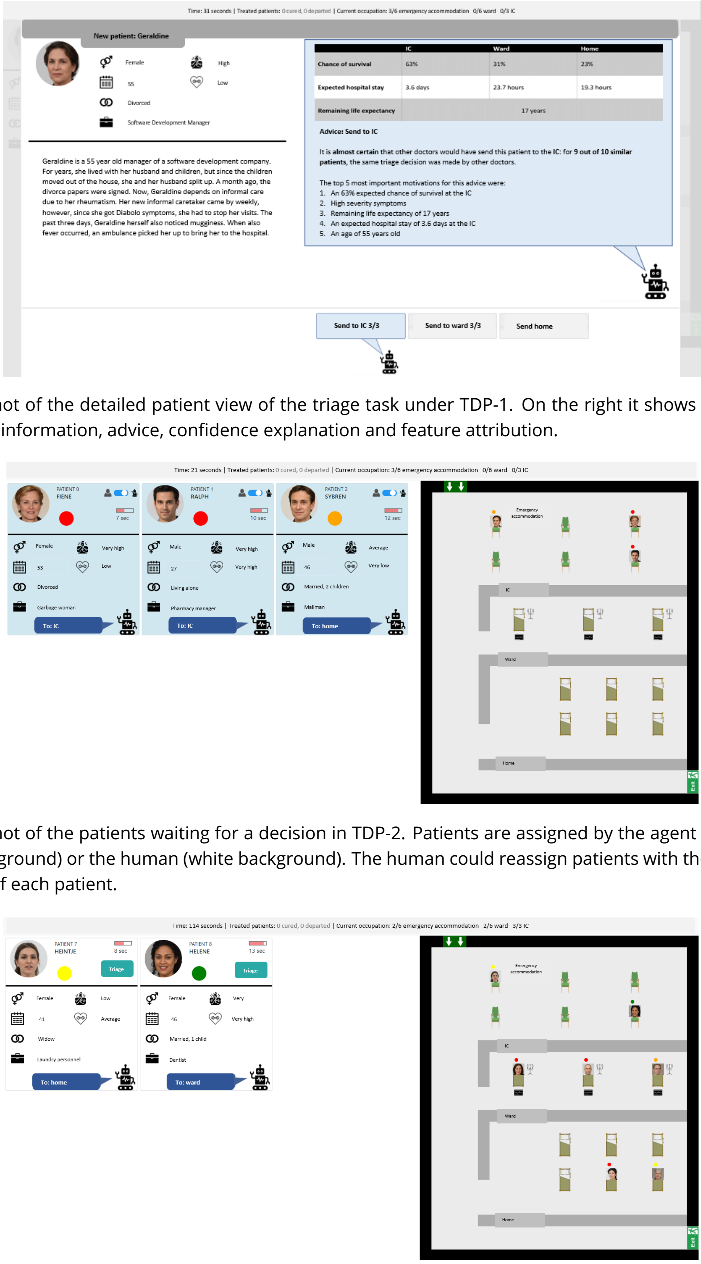
Within the present study, this triage task was performed by a single human and agent although the task allows for larger teams. Furthermore, every TDP resulted in a unique implementation of the agent and interface. The decision support agent of TDP-1 was trained using crowd-sourced labels on several patients and its advice was embedded in the additional patient information (see Figure 8.3a). The task allocation agent of TDP-2 was elicited using a questionnaire about moral values and its allocation was embedded in the patient overview (see Figure 8.3b). Finally, the autonomous agent of TDP-3 used the same elicited values from TDP-2 but performed the task autonomously, waiting a fixed time per patient before enacting a decision (see Figure 8.3c).

In all TDPs, explanations from the respective agents were given on various moments; in the patient overview, the detailed patient view, or when making a decision against the agent's advice or task allocation. The explanation content for TDP-2 en TDP-3 were generated in real-time as they were dependent on the value elicitation outcome. The explanation content for TDP-1 was generated beforehand as the provided advice for a decision were computed beforehand based on the fitted agent.

8.6 The present study

This chapter is concerned with exploring and conceptualizing meaningful human control in human-agent teams by performing early experimentation. We consider a variety of designs for the collaboration between humans and agents, and expect each of them to have different effects on teamwork, and the (experienced) level of human control. The human study serves to obtain evidence for these claims and guide future research. We developed three distinct types of human-agent collaboration in the form of TDPs (see Section 8.4.1), and implemented these in a medical triage task. This forms the basis for our research on Meaningful Human Control. This is a first study, in which we present to healthcare experts our experimental environment, the task to be performed, and the designed human-agent collaborations.

The objective is to investigate how domain experts evaluate: the ecological relevance of the task; the potential value and possible obstacles of agents as their partners in the task, the impact of different TDPs on the control over the team's moral compliance, and the role of explanations to support that control. This first study is therefore qualitative in nature. Results will be used to adjust and improve the current TDPs or create new designs, including the use, presentation and design of the explanations. Results will also be used to improve upon the experimental task, measurements and methods. Further studies can thus better investigate the effects of TDPs on human control in a comparative and quantitative manner.



Time: 31 seconds | Treated patients: 0 cured, 0 departed | Current occupation: 3/6 emergency accommodation 0/6 ward 0/3 IC

New patient: Geraldine

Female

55

Divorced

Software Development Manager

High

Low

Geraldine is a 55 year old manager of a software development company. For years, she lived with her husband and children, but since the children moved out of the house, she and her husband split up. A month ago, the divorce papers were signed. Now, Geraldine depends on informal care due to her rheumatism. Her new informal caretaker came by weekly, however, since she got Diabolo symptoms, she had to stop her visits. The past three days, Geraldine herself also noticed mugginess. When also fever occurred, an ambulance picked her up to bring her to the hospital.

	IC	Ward	Home
Chance of survival	63%	31%	23%
Expected hospital stay	3.6 days	23.7 hours	19.3 hours
Remaining life expectancy	17 years		

Advice: Send to IC

It is **almost certain** that other doctors would have send this patient to the IC: for **9 out of 10 similar patients**, the same triage decision was made by other doctors.

The top 5 most important motivations for this advice were:

1. An 63% expected chance of survival at the IC
2. High severity symptoms
3. Remaining life expectancy of 17 years
4. An expected hospital stay of 3.6 days at the IC
5. An age of 55 years old

(a) A screenshot of the detailed patient view of the triage task under TDP-1. On the right it shows the additional computed information, advice, confidence explanation and feature attribution.

Time: 21 seconds | Treated patients: 0 cured, 0 departed | Current occupation: 3/6 emergency accommodation 0/6 ward 0/3 IC

PATIENT 0
FIENE
7 sec
Female
Very high
53
Low
Divorced
Garbage woman
To: IC

PATIENT 1
RALPH
10 sec
Male
Very high
27
Very high
Living alone
Pharmacy manager
To: IC

PATIENT 2
SYBREN
12 sec
Male
Average
46
Very low
Married, 2 children
Mailman
To: home

(b) A screenshot of the patients waiting for a decision in TDP-2. Patients are assigned by the agent to either itself (blue background) or the human (white background). The human could reassign patients with the slider on the top-right of each patient.

Time: 114 seconds | Treated patients: 0 cured, 0 departed | Current occupation: 2/6 emergency accommodation 2/6 ward 3/3 IC

PATIENT 7
HEINTJE
8 sec
Triage
Female
Low
41
Average
Widow
Laundry personnel
To: home

PATIENT 8
HELENE
13 sec
Triage
Female
Very
46
Very high
Married, 1 child
Dentist
To: ward

(c) A screenshot of TDP-3 where the agent decides for all patients autonomously. Each patient has a time window in which the human could review the feature attribution if so desired.

Figure 8.3: Three screenshots of the three TDPs in the triage task with artificially generated patients.

8.6.1 Agent implementations

For each of the three TDPs, an agent was constructed to let domain experts experience collaboration with such an agent. These were simple rule-based agents to reduce complexity and stochasticity during the human study. The implementation of both the task and agents is publicly available [2].

TDP-1, Data-driven Decision Support, used an agent whose advice was based on crowd sourced data. A total of ten non-experts were confronted with each of the 16 patients and asked to assign care to each with no resource constraints. They were also asked to rate each aspect of a patient (e.g., symptom severity, age, etc.) for their role in their made decision. The decisions were aggregated, to arrive at an ordering of possible care (IC, ward or home) for each patient. During the human study, the agent selected the most frequently selected care if available and otherwise select the next. The ratings were used to manually create the explanation types. For example, the feature attribution explanations container the top-5 of most mentioned patient aspects given that assigned care.

The behaviours of both agents from TDP-2 and TDP-3 were defined by a scoring mechanism to each possible care (IC, ward or home). Given some patient P and our two-part scoring mechanism, we summarize this rule-based decision process as:

$$Care(P) = \begin{cases} \text{IC}, & \text{if } Score(P) \geq 2.5 \\ \text{Ward}, & \text{if } 1.5 \leq Score(P) < 2.5 \\ \text{Home}, & \text{otherwise} \end{cases}$$

$$\text{Where } Score(P) = BaseScore(P) + ElictedScore(P)$$

A patient's base score was defined by its symptom severity;

$$BaseScore(P) = \begin{cases} 3, & \text{if } P_{symptoms} = \text{Severe} \\ 2, & \text{if } P_{symptoms} = \text{Average} \\ 1, & \text{if } P_{symptoms} = \text{Mild} \end{cases}$$

The $ElicitedScore$ was determined using a set of rules obtained from a questionnaire before TDP-2. See Table 8.4 for an overview of these questions. Each question addressed a patient demographic aspect that could influence the decision. Depending on the answers, a rule was selected that added or subtracted 0.25 to the base score. As such, the elicited rules contributed a total of $+1$ or -1

[2] The implementation of both the task and agents is available under the MIT License (2020) and can be found on this link: `https://github.com/matrx-software/MHC-for-triage-agents`.

Demographic	**Answer options**	**Associated rule**
Age	No priority.	-
	Prioritize patients above 60.	If $P_{age} \geq 60$, then $+0.25$ else -0.25.
	Prioritize patients below 60.	If $P_{age} < 60$, then $+0.25$, else -0.25.
Profession	No priority.	-
	Prioritize patients with medical profession.	If $P_{profession} =$ Medical, then $+0.25$ else -0.25.
	Prioritize patients with no medical profession.	If $P_{profession} \neq$ Medical, then $+0.25$ else -0.25.
Gender	No priority.	-
	Prioritize men.	If $P_{gender} =$ Male, then $+0.25$ else -0.25.
	Prioritize women.	If $P_{gender} =$ Female, then $+0.25$ else -0.25.
Family situation	No priority.	-
	Prioritize patients with children.	If $P_{children} = True$, then $+0.25$ else -0.25.
	Prioritize patients without children.	If $P_{children} = False$, then $+0.25$ else -0.25.

Table 8.4: An overview of the elicitation questionnaire, showing the four demographics questioned, the possible answers and the associated rule that adjusted the patient's triage score (a higher score resulted in intensive care).

to the score. In case two patients had the same score and only one bed was available, the TDP-2 agent assigned them to the human and the TDP-3 autonomous agent assigned the care to the first patient.

8.6.2 Methods

8.6.2.1 Design

Team Design Pattern is manipulated within-subjects. All our participants were domain experts and practiced the triage task under each of the three TDPs, in the following order: solo, without the agent being involved (baseline); with the agent providing decision advice (TDP-1); with dynamic task allocation between human and agent (TDP-2); and with the agent acting autonomously according to a model of the human's moral values (TDP-3).

8.6.2.2 Recruited domain experts

A total of 7 health care professionals participated in this human study. Four of them have a history as volunteers in health care to conduct triage (e.g., volunteer of the Red Cross); three worked in a hospital as medical professionals. The latter were more experienced in working with intelligent machines.

The human study was approved by the TNO ethics committee. The domain experts were recruited through personal and professional networks. Inclusion criteria were an academic background and experience in the healthcare domain. All experts stated to have sufficient technical ability to participate in an human study held in a digital environment. A €25,- compensation and travel reimbursement was offered. One expert did not perform the triage task with TDP-2 due to technical issues.

8.6.2.3 Measures

The objective of the present study is to obtain information how domain experts appreciate and evaluate the distinguished designs of collaboration with intelligent agents when performing a morally sensitive task. We are particularly interested in how the experts assess the control they experience over the task processes and the decision making, and whether this differs for the distinguished designs of human-agent collaboration. Furthermore, the study aims to obtain information how the experts understand and evaluate the explanations provided by the agents, and whether they consider these explanations as supportive for the collaboration.

In order to obtain the participating expert's assessment, thinking-aloud and semi-structured interviews methods were used. The experts were asked to think aloud while they were carrying out the task. Afterwards the experimenter asked them questions with respect to their experiences and opinions. In order to obtain input for these interviews, a series of exercises and questionnaires were administered. Below we provide a concise description of the measurements used.

Semi-structured interviews

The key measure used was that of a semi-structured interview, to which the other measures provided input. The nature of the interview was interactive and open. The goal of the questions was to guide the experimenter during the conversation and to collect qualitative data. As such, the proposed questionnaires discussed below were by no means intended as stand-alone justified measures. Responses to these questions were used to ask open-ended questions to acquire a free-format and detailed account of the domain experts' experiences.

The interview started with questions about their profession and experience with (intelligent) machines, followed by questions regarding the collaboration, control and explanations. For instance, questions were asked regarding their preferred TDP and motivation for this preference. When necessary, the experimenter could ask follow-up questions. An example of this was to investigate why a particular expert was optimistic about HAT solutions for the health care domain (e.g. "Why are you positive about the collaboration between human and machine in the health care domain?").

Thinking Aloud

The interviewed experts were instructed to think aloud when they were performing the task [399], especially concerning how they experienced the collaboration with the agent. If needed, the experimenter prompted the expert to not only describe *what* they were doing, but also to verbalize the *why* of their thoughts and actions.

Ecological validity

In order to reveal how the healthcare professionals evaluated the ecological representatives and validity of the task, we administered two written questions (translated from Dutch): (1) "From 0 to 100, to what extend does our interpretation of medical triage match yours?", and (2) "From 0 to 100, to what extend do the induced stressors match with what you expect in reality?". After scoring each, a brief open interview with the experimenter followed regarding their scores. These questions assessed the experts' judgment on: (1) the provided information and the administered triage task, and (2) the introduced task stressors, such as the induced time pressure and the imposed limitation of available resources.

Control

Unfortunately, standardized and validated questionnaires for measuring a participants' control over task performance in a human-agent context do not yet exist (see Section 8.2.2). We therefore composed such a questionnaire, consisting of eight statements (see Table 8.5). For each statement, the interviewed experts were asked to indicate their level of agreement on a 5-point Likert scale. The goal of this questionnaire was to identify the expert's initial experiences, providing input for relevant follow-up questions regarding their answers.

		Strongly disagree		-		Strongly agree
1.	It was difficult to keep an overview of patients and available resources.	1	2	3	4	5
2.	I experienced time pressure during decision making.	1	2	3	4	5
3.	I felt responsible for the well-being of patients.	1	2	3	4	5
4.	I made decisions under inconclusive medical- and ethical guidelines.	1	2	3	4	5
5.	I made decisions during the task that I would not want to make in real life.	1	2	3	4	5
6.	I felt uncomfortable during (some) decisions I made.	1	2	3	4	5
7.	I mostly made decisions for patients that led to a good division of care.	1	2	3	4	5
8.	I mostly made decisions that led to a good division of care for all patients.	1	2	3	4	5

Table 8.5: The statements used to serve as input to the semi-structured interviews about the experienced control.

		Strongly disagree		-		Strongly agree
1.	I found that the data included all relevant known causal factors with sufficient precision and granularity.	1	2	3	4	5
2.	I understood the explanations within the context of my work.	1	2	3	4	5
3.	I could change the level of detail on demand.	1	2	3	4	5
4.	I did not need support to understand the explanations.	1	2	3	4	5
5.	I found the explanations helped me to understand causality.	1	2	3	4	5
6.	I was able to use the explanations with my knowledge base.	1	2	3	4	5
7.	I did not find inconsistencies between explanations.	1	2	3	4	5
8.	I think that most people would learn to understand the explanations very quickly.	1	2	3	4	5
9.	I did not need more references in the explanations: e.g., medical guidelines, regulations.	1	2	3	4	5
10.	I received the explanations in a timely manner.	1	2	3	4	5

Table 8.6: An adapted form of the System Causability Scale by Holzinger et. al [400] to provide input on the semi-structured interview on the quality of the offered explanations to support control.

Explanations

To obtain information from the domain experts as to how they appreciated the provided explanations, and how they valued the role of these explanations for their collaboration with the agent, an adapted version of the System Causability Scale (SCS) [400] was administered after completing each round. The adapted SCS consisted of ten questions (see Table 8.6). Again, these answers were used to ask detailed follow-up questions to explore the experts' experiences with the collaboration.

Usefulness of explanation types

Each TDP utilized one or more of the three explanation types as discussed in Section 8.3. In order to gain insights in the perceived usefulness of these dif-

		Strongly disagree		-	Strongly agree	
1.	The explanations helped me during task performance.	1	2	3	4	5
2.	The explanations mostly confirmed me in what I already knew.	1	2	3	4	5
3.	The explanations provided me new information.	1	2	3	4	5
4.	The explanations led to new insights.	1	2	3	4	5
5.	I understood the explanations well.	1	2	3	4	5
6.	The explanations helped me to determine whether I could trust the computer.	1	2	3	4	5
7.	The explanations made me reason about how to make triage decisions.	1	2	3	4	5
8.	The explanations gave me new insights of how intelligent systems should support humans.	1	2	3	4	5

Table 8.7: The statements used to provide input on the semi-structured interviews regarding the perceived usefulness of the explanations. These were provided together with a screenshot of a single explanation type used in that condition.

ferent types, screenshots of the explanations were presented and seven statements were provided (see Table 8.7). Each expert was asked to indicate its level of agreement using a 5-points Likert scale. These statements were developed as an extra and more systematic approach, next to the semi-structured interview questions, to gain valuable insights in the explanation types. It would evoke follow-up questions for the semi-structured interview. For example; "What in the particular explanation helped you gain trust in the intelligent system?".

8.6.2.4 Procedure

The participating domain experts took part on a one-to-one basis (one expert, one interviewer). A session took approximately two hours, held in November 2020 within the Netherlands. First, the experimenter explained the goal and nature of the study, and provided an outline of the procedure. The expert read the information sheet and signed the informed consent form.

The expert received a detailed instruction to the triage task and were instructed to read the scenario of the pandemic, as well as the ethical and medical guidelines to triage for the present study (which were based upon actual Dutch guidelines). Here, the expert was also motivated to ask clarifying questions at any time.

Then, the expert was asked to conduct triage in the implemented testbed without the help of an artificial agent. In this baseline condition, 16 patients had to be triaged. The expert was instructed to apply the given ethical and medical guidelines. After completion of the baseline task, questions addressing the ecological validity were administered (as proposed in Section 8.6.2.3).

The expert was then asked to triage a new set of 16 patients, this time receiving decision advice from their personal artificial agent according to TDP-1. During the task, instructions were given to think aloud (TA), which was recorded while notes were taken. When the expert considered the guidelines to be indecisive or inappropriate, instruction was given to follow their own personal moral values and decide accordingly. After completion of the task, subjective measurements were taken concerning *experienced control* and the *value* and *usefulness of explanations* to serve as input for the semi-structured interview afterwards. This process was repeated for TDP-2 and TDP-3.

When all conditions were completed, the expert was asked to reflect on all three TDPs. An indication had to be given which TDP they would prefer to use in their work (if at all), and why. After the interviews, all collected data was anonymized using pseudonymization and a key-file that was removed after two weeks.

8.6.3 Results

The findings per TDP will be reported as follows:

- **Team collaboration**: The participating expert and the agent jointly performed the assigned task of assigning medical care to a set of patients. The nature of this human-agent collaboration was shaped by the particular TDP. Per TDP we report on how the expert evaluated the collaboration and the task division.

- **Control**: Per TDP we report if the domain experts experienced to be in sufficient control to ensure that all decisions were made according to their own personal moral values. For this, we used results from questionnaires and interviews, seeking for trends in how much control the expert experienced.

- **Explanations**: For humans to exercise control efficiently and accurate, they need to have an understanding about their agent partners which explanations can help ascertain. Outcomes from the explanation questionnaires and semi-structured interviews were used to report how the experts evaluated the agent’s explanations, and if they supported a better understanding of the team.

8.6.3.1 Ecological validity

The domain experts scored the ecological validity of the used scenario in which medical triage was conducted and the available information for doing so with an average of 75.71 (*SD*=10.50*)* out of 100. They scored the ecological validity of the stressors in the task with 76.43 (*SD*=6.39*)* out of 100. Two respondents indicated that the time pressure induced by the pace of patients being submitted for triage was too high. We take these findings as a reassurance that the developed testbed and task is suitable for investigating MHC in a situated manner.

8.6.3.2 Findings TDP-1

In this design of human-agent team collaboration, the agent provides information and gives advice with the human making the actual triage decision.

Team collaboration

The domain experts evaluated this pattern of collaboration with the agent fairly positively. Three out of the seven experts preferred TDP-1 over the other two Team Design Patterns. Two out of the seven experts mentioned TDP-1 in combination with TDP-2 as their ideal collaboration with an intelligent system. Furthermore, they pointed out that the agent did what computers are best at, discovering and presenting statistical relationships in the domain; and that they themselves could concentrate on making decisions. One interviewee said: “The agent provided quick computational power to calculate valuable data, whereas I as a human could make the actual moral decision".

The interviewer asked each expert how they experienced the role of the decision support agent. They indicated that if the agent’s advice corresponded to their initial opinion, the congruence was regarded as a confirmation that it was an appropriate decision. If, however, the agent’s advice deviated from the own opinion, then this was for many the sign to change their decision. Overall the experts elaborated that they interpreted the agent’s advice to be representative for what doctors in general decide. One expert argued: “all those other doctors probably know best".

Control

The domain experts found the task with the decision support difficult and strenuous. Most experts pointed out during the interview to feel responsible to assign the best possible care to all patients, which aligned with the results of question 3 of Table 8.5. Four experts scored a “totally agree" on the experienced responsibility over the patients well-being. The other three experts scored this with a “very much agree". They said to realize that a swift processing was important, as to prevent deterioration of a patient’s condition pending their triage decision, subsequently experiencing stress about this. On being asked whether they judged this as a threat to maintaining control, most experts indicated to be able to cope with the time pressure. They said that the support offered by the agent (such as computational information about a patient's survival chances for every possible care) helped them to manage the imposed time pressure. Overall, the experts reported to experience adequate control over the process and decision making. Furthermore, all evaluated their triage decisions to be compliant with their own moral values and the provided ethical guidelines.

Explanations

The predominant response of domain experts was that the conditions imposed by the experimental simulation did not allow them to form a proper judgment about the value of the explanations. They felt to be working under extreme time pressure (see above), which precluded them to process the explanations. One expert remarked: "all that text took too much time to read", and suggested to provide explanations in a visual form instead. Another expert indicated that the assumption that "the agent acted as a representative of other human doctors", allowed him to disregard the explanation altogether. Ironically, one purpose of explanations in human-agent teams is to establish and support *appropriate trust* in each other. Thus, during the interview the experts mentioned to be unable to conduct a proper evaluation of the given explanations. However, in response to the questions about the value of explanations, they rated the explanations as neutral to positive for achieving a better understanding. Here, an average of 3.38 (*SD=0.49)* was given to the usefulness of explanation types where 1 was considered as "Strongly disagree" and 5 as "Strongly agree" (see Section 8.6.2.3).

8.6.3.3 Findings TDP-2

In this design of human-agent team collaboration, the agent proposes which patients to assign to the human and which patients to the agent. This is done based on an earlier value elicitation process using a questionnaire. The agent can provide an explanation for its intended decision as well as the allocation of each single patient. The human can overrule this allocation. Each patient is independently triaged by both human and agent based on this (overruled) allocation.

Team collaboration

The interviewer asked the domain experts how they experienced the role of the dynamic task allocation agent. They indicated that the division within the triage task helped them focus on their own patients. Also, they experienced the task to go *faster* in comparison to TDP-1, which was evaluated as a pleasant effect. Two experts mentioned that reviewing the explanation took too long and would have a detrimental effect on task performance. Instead, they kept patients allocated to the agents without reviewing the relevant explanations.

All experts mentioned to *trust* the agent in the decisions it made for the patients that were assigned to it. The overall motivation was that they understood and accepted why and how the agent assigned patients. To quote one expert; "I understood why the intelligent system assigned certain patients to itself (...), and that its decisions were based on my value elicitation".

Control

During the interview most domain experts indicated that they considered it a challenge to maintain an overview of the patients requiring a triage decision. One expert rated complete agreement (5/5), and four a strong agreement (4/5) on question 1 of Table 8.5 of not being able to keep an overview. These expert argued that they felt a need to continuously monitor all patients, including those assigned to the agent. This required too much effort according to them, to exercise adequate control. When asked to elaborate on this, they argued that the agent's triage decisions (e.g., assigning a patient to the IC) had an impact on their own decision space (e.g., all available IC-beds occupied). Keeping overview on what the agent was doing, while simultaneously paying attention to their own patients often imposed too much pressure to exercise adequate control, as three experts emphasized explicitly.

Opposed to the five experts experiencing high workload, the other two experts reported to not feel this pressure. They explicitly reported to rely on the qualities of the agent and its triage decisions. When asked what caused this reliance, they argued that they had noticed the agent to comply to the their personal moral values, assessed earlier during elicitation. This resulted in a feeling for them that decisions could be safely dealt with by the agent. Which in turn made the experts experience more time available to focus on assigning care to their own patients.

On average the experts felt slightly less responsible for the patient's well-being compare to TDP-1 (TDP-1 scored an average of 4.58 (*(SD=0.49)* and TDP-2 scored an average of 4.16 *(SD=0.68) on question 3 from Table 8.5*).

All experts were positive about the option to overrule the agent's assignment. This was evaluated as a valuable asset of this TDP and they all indicated that it contributed to their experienced control.

Explanations

Five of the experts mentioned that they *missed* the statistical data that was presented in TDP-1. They interpreted this data as an explanation of agent reasoning, even though it was not presented as such.

Similar to the previous condition, the view on how the explanations were presented was referenced by all experts. Again, it was suggested that visualizations might be beneficial, since the provided text took them too much time and effort to interpret.

Overall, the explanations were not utilized excessively, as four experts reported. However when they felt they had the time, they were perceived as helpful, establishing a form of *trust* and *understanding* of the system. The open-ended interview question on whether the experts considered the explanations as important, all answered with “yes". One expert indicated that: “The explanations help me during the task. If these would not be provided, it would have been very unpleasant".

8.6.3.4 Findings TDP-3

In this design of a human-agent team collaboration, the agent autonomously makes all decisions swiftly based on the elicited moral values elicited before the task. The human observes the agent making these decisions to understand how the elicited values impact agent behavior and to make adjustments next time if needed. More information about the agent’s reasoning behind a decision could be requested. Note that this collaboration does not allow the human to exercise instantaneous control such as intervening in an agent’s decision.

Team Collaboration

All experts reported this collaboration as uncomfortable during the interview. Two explicitly motivated this by the fast pace patients entered the environment, and two with not being able to overrule the agent. In some cases, experts were not motivated to request an explanation on why certain decisions were made. One argued: “I do not feel part of a team, because I don’t play a role in the decision making process". As a result, the experts did not feel responsible for the decisions made by the agent, similar as in TDP-2.

Control

In all cases, the experts stressed the discomfort that arose from not having the opportunity to overrule the intelligent agent. When asked about their *trust* in the agent, two experts responded that the agent was compliant to the earlier given value elicitation. Three also mentioned they understood the reasoning of the agent, which also established trust. Interestingly, one mentioned that this did not meant that (s)he always agreed with the triage decisions made by the agent.

The four experts who noted a high pace and time pressure, also lacked motivation to take on a supervisory role in the collaboration. Their reason was the stress and lack of overview evoked by this pace.

Explanations

The two experts who reported on the uncomfortable high pace, indicated to seldomly read the explanations. One of them commented: “If I read one explanation, I miss out on three other patients and the decision made for those.". The five experts who indicated they did read the explanations, scored on average 3.67 (SD=0.75) to question 5 of Table 8.6. Indicating they found the explanations useful in understanding the agent’s reasoning.

The interviewer asked these five experts about the trigger for wanting more information about the agent’s reasoning. Two explained that this was only when they did not agree with the decision made by the agent. They expressed a curiosity in *why* the agent would make such a decision to be able to better understand the system. The other three explained to have an overall curiosity, independent of the made decision.

8.6.3.5 Comparative findings

This section highlights similarities and differences within the three collaboration designs (summarized in Table 8.8):

Team Collaboration

Overall, all interviewed domain experts reported TDP-1 as their preferred collaboration design. They substantiated this preference by the clear division between human and machine, which was appreciated.

Notably is that four out of seven experts motivated their interest in TDP-2, especially when the agent would be “more mature", as one expert described it. When asked to elaborate on this, they referred to the agent in TDP-1 who provided additional information as well as advice. Effectively, they proposed a combination of the data-driven decision support agent from TDP-1 with the dynamic task allocation functionality of the agent in TDP-2, implying that the data-driven agent would make its own decisions.

In all three conditions, the speed of incoming patients was emphasized. Interestingly, in TDP-1 and TDP-3 this was perceived as unpleasant, while in TDP-2

Aspect	General findings	TDP-1	TDP-2	TDP-3
Team collaboration	Valued the more direct control was experienced.	Most valued due the direct control.	Mostly valued for its high potential.	Did not feel like a collaboration.
Sense of control	Only when capable of influencing decisions directly	High degree due to agent not making any decisions.	Some experienced control, due to feeling capable of intervention.	No feeling of control, as it was a delayed form of control.
Use of explanations	Perceived as useful in hindsight, but not actually used.	Intermediate statistics were found most valuable.	Explanations were not used or ignored.	Observing behaviour more useful than explanations due to agent decision speed.

Table 8.8: A summary of the key findings separated on the three aspects measured and the Team Design Patterns tested. In TDP-1 the agent provides advice to an expert making decisions. In TDP-2 the agent distributes decisions between itself and the expert, which can be overwritten. In TDP-3 the agent made all decisions under supervision according to elicited decision rules beforehand.

the speed of assigning patients as a team was being appreciated. In fact, in this condition one expert mentioned to deliberately leave patients assigned to the agent, as to make sure those patients received care earlier.

Furthermore, there was a sense of confirmation amongst all experts in TDP-1 and TDP-2. They experienced it as helpful when the advice (TDP-1) or decision (TDP-2) was congruent with their own initial decisions. Within TDP-1, they felt supported by the agent rather than collaborating with it. As such, TDP-1 did not result in a feeling of collaborating with the agent. The supervised autonomy collaboration from TDP-3 received the least willingness from experts to collaborate, since they felt they could not take part in the decision making process.

Control

A comparison of the results per TDP revealed that TDP-1 was favoured by three experts when it came to the *sense of control*. Furthermore, when the interviewer asked to rank their top 3 of the TDPs, TDP-1 was placed first by five experts. The main reason for this was their sense of having complete control over decision making. In contrast to TDP-1, TDP-3 was the least preferred by six experts due to the inability to directly influence the decision making process. The seventh expert who favored this TDP, only did so under the condition that the human team member would receive the ability to overrule the agent.

Even though *speed* was mentioned to be an asset in TDP-2, it also effectuated a lack of overview. Because both expert and agent could influence the environment, all experts had difficulties keeping track of the situation. However, they did experienced control as they could influence on the task allocation. In other TDPs, increased collaboration speed resulted in increased time pressure, overall resulting in experiencing less control.

Furthermore, none of the experts experienced the ability to exercise control through the value elicitation process, which determined the agent's decision making in TDP-2 and TDP-3. At times, they mentioned that they felt the agent did comply to the elicited moral values, but this did not result in an experience of being in control. This was especially the case in TDP-3, where intervention was not possible as was the case in TDP-2.

Lastly, TDP-3 evoked the least feeling of responsibility over patients in comparison with TDP-1 and TDP-2. The lack of experiencing control through value elicitation over the agent negatively impacted the sense of responsibility for the agent's decisions.

Explanations

In general, explanations were found useful, as all experts mentioned in the semi-structured interview, but mostly in retrospect as they were utilized only occasionally in TDPs. The main reason for not requesting an explanation was the experienced time pressure.

The most common trigger for requesting an explanation from the agent was when the agent showed incongruency with the expert's initial own decision. In those cases, it established a form of *trust* as well as better *comprehension* of the system according to the experts.

During TDP-3, all experts felt overwhelmed by the agent's decision speed and felt they learned more from observing the agent than by reading explanations.

8.7 Discussion

The advancements of embedded artificial intelligence allows modern technology to conduct complex tasks more and more. This provides new opportunities, but it also raises the question whether and how humans can still exert meaningful control over the technology's behavior. This is especially important for tasks for which ethical and moral values apply. To address this question it is important to first define multiple manners of organizing the contribution of humans and technology in human-agent teams. Subsequently, research is needed into how these different patterns of human-agent collaboration affect the human's control over the agent's behavior and the team's performance. This chapter addressed both needs.

Three different design patterns for human-agent collaboration have been developed and implemented into a medical triage task. Then, medical domain experts were asked to work with these agents under these team design patterns. Multiple qualitative methods and measures were used to learn how the domain experts experienced the collaboration with the agents, and, in particular, how well they felt in control over the task and over the decisions taken by the team. We feel that this kind of qualitative research is important to obtain a better understanding of how different collaboration options affect the feasibility of humans to exert meaningful human control. This understanding is needed to define and refine the team design patterns for use in future practical applications.

Findings indicated that the domain experts wished to make as many decisions as possible even when experiencing an already high workload. Furthermore, they only felt responsible for their own decisions instead of all decisions made within the team. As such we identify several challenges in the design of a human-agent team based on these finding. First, the way humans and agents collaborate should ensure that humans feel responsible for agent behaviour, otherwise they will lack the motivation to exercise the necessary control. Secondly, humans need to be prevented from exercising control unnecessary often when they do feel responsible as this will increase their perceived workload. This includes protecting humans from their tendency to take on too many decisions. These two challenges could be solved by making agents more aware of the human's mental state to adjust the way they collaborate (e.g. through physiological measurements or lack of response time thresholds).

Several findings indicated that the experience of control depends on how immediate its observed effects were. Two control mechanisms were evaluated varying in how instantaneous their effects were, with the domain experts favouring the more instantaneous control (task reallocation) over the other (iterative value elicitation). Depending on the task, instantaneous control is not possible or not doable for humans. When a effects are delayed, agents could explain the consequences of the exercised control to humans. This 'consequential explanation type' could improve the experienced control as well as support the human assessment if the exercised control would have the intended effects. These type of explanations could for instance include simulations of agent behaviour given the intended control signal.

Different types of explanations were used in the various collaboration designs. Their purpose was to improve control and calibrate trust by enabling an accurate mental model of the agent's reasoning. However, findings indicated that the experts almost never reviewed the explanations due to the experienced time pressure caused by their desire to decide quickly. Interestingly they did value all the explanations in retrospect and could see their potential. This shows that research into agent-generated explanations should not only focus on their content but put equal, if not more, focus on how and when they are presented. Ideally, the agent should be made aware of human workload and adjust its explanations accordingly. This again underlines the need for agents to be aware of the human's mental state, not only to adjust how they collaborate but also how they explain.

Some findings indicated that the domain experts trusted the agents too much, and relied at times more on the agent's judgement than their own. Even though they were instructed to be critical as they were held responsible for the agent's behaviour. This indicates a potential automation bias. This would explain why they did not feel the need to make time to review and interpret the explanations. Interestingly, the explanations were intended, among others, to counter this over trust. Within the field of Explainable AI it is often mentioned that explanations support appropriate trust calibration. However, if too much trust prevents humans from interpreting the explanations, those explanations become meaningless. Further research on the relation between automation bias and the use of explanations is warranted. If this hypothesis proves to be true, agents need to be aware how much their human partners rely on them and adjust their way of presenting explanations accordingly.

A limitation of the study is its qualitative nature to evaluate the different collaboration designs. Although, as we argued, a qualitative study is better suited in exploratory research as it provides more information compared to a quantitative study. However, only one task was used with specific properties, making it difficult to generalize the results. Future studies should focus on further similar evaluations with tasks combining both objective and subjective measurements. Research could expand our designs to include agents who model and track the human mental state and adjusts their collaboration and explanations accordingly. Specifically, further study is warranted on how agents can foster a human feeling of responsibility and facilitate the experience of control when its effects are delayed. Finally, it is important to research how agents should formulate and present their explanations such that humans feel they have the time and need to review them.

8.8 Conclusion

This chapter addressed the design of human-agent collaborations, specifically in morally sensitive domains where humans should have meaningful control over the agents. This control needs to ensure that team behaviour is compliant to human moral values and ethical guidelines. Otherwise, the human should be able to be held responsible. Three of such collaboration designs were presented. Each varied in the agent's autonomy and we evaluated the experienced control and the value of provided explanations in each using structured-interviews with domain experts.

These three design patterns and the performed interviews form a first iteration towards designing human-agent teams that support meaningful human control for various tasks. A design pattern approach was taken, and a first set of measurements were introduced together with a reusable testbed to evaluate human-agent collaboration to support meaningful human control.

Results from the expert interviews showed that the used task of medical triage was sufficiently realistic and its simulation valid. Furthermore, we found that how responsible humans feel for agent decisions relates to their involvement in those decisions. If the experts only supervised they did not feel responsible, even though they could exercise control by defining agent behaviour beforehand. The more the experts felt in collaboration with agents, the more they felt in control. With having sufficient time and influence over the agent as prerequisites for this collaboration. To support this, agents could benefit from monitoring the workload of humans and adjust their collaboration form and explanations accordingly. Specifically when humans experience time pressure, agents could motivate humans to remain involved in their decisions as they can be held responsible for them. In addition, agents need to adjust when and how they communicate their explanations in these cases to also motivate humans to review their explanations. As the results from our interviews showed that pressured humans might tend to trust the agent too much and ignore its explanations designed to prevent such over-trust.

This chapter presented a first step in exploring how domain experts experience human-agent team designs that aim to enable meaningful human control supported by explanations from the AI.

Acknowledgments

This research was funded and conducted in two TNO projects; RVO Human–AI Teaming and ERP-AI Man-Machine Teaming. We thank the healthcare professionals for their participation in our study and their organisations for their approval.

PART IV

CONCLUSIONS AND DISCUSSION

CHAPTER 9

DISCUSSION

In this thesis, we aimed to design and develop explanations an AI agent can and should provide to support a responsible and effective collaboration with humans. In this chapter we review our main findings towards this aim and discuss them with respect to the literature. First, we discuss our findings from Part I where we addressed the contrastive explanation class to disclose how an AI agent arrives at a decision. Next, we discuss our proposed confidence and actionable explanation classes from Part II that support respectively trust calibration and contestability in an AI agent's decision. We then discuss our work presented in Part III on how design patterns can be extended and used to embed explanations in the design of human–AI collaborations.

After discussing our main findings and reviewing each posed research question, we discuss the limitations of our research. These relate to our performed user studies, benchmark tests, the post-hoc and one-shot nature of our explanations and the model-agnostic approach of methods proposed. We also review the gaps in our research, in particular that we did not research the modality of explanations which is expected to have an impact on several of our conclusions.

Finally, we reflect on the societal implications of our work. We do so based on current societal developments with respect to upcoming regulations and examples of (unwitting) misuse or bad design of AI agents, as well as based on our own experience and conversation with the industry, government institutes and relevant organizations who are looking to adopt XAI research results in their AI agents.

9.1 Findings: Part I

In Part I, we began with addressing how an AI agent should explain the reasons for a made decision. We addressed the contrastive explanation class that aims to explain why one decision was made instead of another. Its purpose is to convey a general understanding of the AI agent's decision making in a comprehensible and intuitive format. However, no suitable form to convey such information in an explanation was yet determined [42], hence the question in this part was as follows:

RQ 1: How should an AI agent explain why it made one decision instead of another?

Chapter 2 evaluated two forms of a contrastive explanation in a human study: the rule-based and example-based forms. Both aimed to explain the decision boundary between a current decision and another decision of interest. These are two common forms to convey explanations [57, 45]. We found that only the rule-based form improved the actual understanding, albeit slightly. Both forms did, however, improve our participants' subjective feeling that they understood the AI agent better, although that feeling did not correlate with their actual understanding. In addition, both forms but the the example-based form in particular, resulted in a more persuasive AI agent causing a blind following of the AI agent's advice.

These results indicate that the effects of explanations can be diverse and that not all effects are equally desirable. Depending on the use case, it can become beneficial or detrimental for an AI agent to become more persuasive. For instance, such an effect is beneficial when a persuasive AI agent is required to support a patient in making the necessary behavioural changes to improve their health [401]. However, in other cases creating a more persuasive AI agent is ambivalent, such as the use of AI agents and explanations to protect users in social networks to disclose too much personal information [402]. Here, it is argued that a more persuasive AI agent protects the privacy of users but at the cost of limiting their autonomy. In other use cases, especially those we associate with high risks such as medical triage, we feel a persuasive AI agent is entirely unethical [403]. These examples show the value of our evaluation of contrastive explanations, as our findings, combined with literature on the use of persuasive AI agents, indicate that the application of contrastive explanations should be carefully considered given the use case and its context.

In addition, our results indicate that simply providing information about how an AI agent decides does not necessarily result in an understanding that improves the human–AI collaboration. Interviews with participants revealed that the rule- and example-based explanations felt to them as a collection of independent facts lacking an underlying common rationale. Further questioning revealed that this hindered the participant's ability to recall these facts due to the difficulty of distilling a mental model about the AI agent from the explanations. This would explain the limited understanding they gained from the explanations, even though it caused a feeling of understanding. As they might now feel that they know many facts about the AI agent's functioning, but lack the ability to generalize from such facts. This is the first empirical indication of the theory that explanations can foster an illusion of understanding in AI agents [404]. The instigation of this illusion through explanations creates a profound risk that an

AI agent will be used irresponsibly when it can explain itself only superficially. This risk is further increased when explanations are personalized based on self-reported and observed human cues. For instance, the proposed methods from Kouki et. al [405], Martijn et. al [406] and Sreedharan et. al [407] utilize such cues to generate explanations to which humans react positively. Since the illusion of understanding can be entirely unconscious, such personalization increases the risk of instigating the illusion of understanding through a feedback-loop. A loop that would result in humans feeling a sense of understanding and acting upon that feeling, without having an actual understanding of the AI agent.

Our findings help illustrate the still open challenges in the research towards contrastive explanations for AI agents. The results from this human study show that an explanation can do more than offer a general understanding in how an AI agent makes its decisions. Similarly, the results also indicate potential harmful or unwanted effects an explanation may cause which warrant further studies about such effects. Three recommendations were offered on how to approach such additional studies: 1) offer motivated hypothesized relations between effects and explanation including potential negative effects; 2) decide on a task and population sample that matches the purpose of the explanation and the study; and 3) select multiple appropriate measurements per effect and make use of measurement triangulation to assess the study's validity and credibility. These recommendations assist in the design of rigorous human studies to evaluate an explanation and all the effects it may cause. Without such evaluations, we risk the use of explanations with adverse side effects, without ever achieving the intended effect.

We next addressed how an AI agent could generate the promising rule-based form of contrastive explanations. Having found the rich effects such explanations could cause, we considered whether an AI agent can indeed generate them. Chapter 3 addressed this question for AI agents providing decision support based on classification models. Chapter 4 addressed the question for AI agents based on reinforcement learning who determine their behaviour over time given on a goal to achieve.

Both chapters introduced the idea of using surrogate models to describe the decision-making process of an AI agent. This approach is opposed to the development of an intrinsically interpretable decision-making process [46]. In both chapters, an algorithm was devised where a surrogate model perturbed the AI agent's behaviour, assuming access to only its inputs and outputs. These perturbations were used to extract information on that AI agent's behaviour, forming the basis of the eventual explanation after a computed output. The use of input-output perturbations enabled the development of a method that could generate explanations without setting any requirements on the implementation of the AI agent. This makes the proposed methods generally applicable, both for current and future AI agents. This approach is referred to as the modal-agnostic approach [140].

In Chapter 3 we proposed a method that could generate rule-based contrastive explanations following the model-agnostic approach. We showed this approach to generate these explanations in reasonable time while being faithful to the functioning of diverse classification models and task complexities. With this method, we were able to create AI agents that can explain why they made a decision instead of another by conveying interpretable decision rules that accurately describe the AI agent's behaviour. This method thus describes an AI agent's decision boundary for a made decision. Whereas previous methods were limited to indicate only the most important data features used to make that decision [59, 42].

Chapter 4 discussed how the idea of surrogate models could be extended to generate example-based contrastive explanations for reinforcement learning agents. Only few methods exist that allow an AI agent taught through reinforcement learning to explain itself [408, 409]. There is also little known about what kind of explanation humans expect from such an AI agent [410]. The cause could be that this research community tends to disregard the human who interacts with an – on reinforcement learning based – AI agent and for whom the explanation is intended [411]. To remedy this, we conducted a human study to explore what kind of explanations humans would find desirable from such agents. The results showed that humans prefer explanations about the long-term elements of the devised plan and the consequences the AI agent expects when performing this plan. We hypothesized that this is due to humans wanting to know what the AI agent will do, what it is expecting to achieve and whether it correctly understands how its actions impact the world. Based on these findings, we developed a model-agnostic method to generate such explanations. This method could explain an AI agent's plan in interpretable and abstracted terms as well as answer queries probing why the AI agent opted for its current plan instead of another. Since then, several methods extended our proposed method and findings. See for example Lin et. al [412] and Sreedharan et. al [413] who applied our idea of representing series of actions as abstracted and interpretable concepts or Madumal et. al [414] who extended our idea of consequences with the concept of causal relations between actions and changes.

To conclude this first part, these findings show that it is feasible to generate contrastive explanations for various AI techniques. In addition, contrastive explanations have value as they cause several effects. However, these effects can also be potentially negative depending on the context in which they are used. Additional sound and rigorous human studies are needed to assess such effects in relation to this context. Three recommendations were made on how to conduct such rigorous evaluations.

9.2 Findings: Part II

In Part II we defined two novel explanation classes whose purpose is more specific than simply providing an understanding of the AI agent's functioning. As humans have a specific goal in mind when collaborating with an AI agent, an explanation should support that goal. For example, humans might want to make decisions that are as accurate as possible (e.g., a doctor making a diagnosis). An AI agent can offer advice and support in this pursuit. Another example is when humans need something that is determined (partly) by an AI agent (e.g., the acceptance of a loan). In both cases, it is important for the human to determine when and how to act upon the AI agent's output, whether this is an advice or a decision. In a decision-support setting, the human needs to determine when to trust and rely upon the AI agent's decision (i.e., whether the AI agent is trustworthy) [415]. In a setting where the human is subjected to the AI agent's decision in some way, they want to be able to enact a degree of control over the AI agent or otherwise be able to contest it [416], particularly when that decision is unfavourable to them, limits their autonomy, conflicts with their values, or causes harm in any way [241].

In Part II we thus addressed the following question:

RQ 2: Which classes of explanations enable humans to decide whether and how to act on an AI agent's decision?

Chapter 5 presented the novel class of confidence explanations. These explanations regard how likely it is that the AI agent's advice will prove correct and how this confidence was computed. The purpose of confidence explanations is to assist a human in determining whether its output should be trusted. This explanation class aimed to fulfil the research gap for explanations that explicitly support the human's decision to rely on an AI agent's decision or not [417].

Based on the literature and on two human studies, four properties were defined and validated to define effective confidence explanations; a confidence estimate should be accurate, predictable, transparent, and explainable. Case-based reasoning was proposed to compute confidences matching all four. This approach defines confidence as the likelihood of a correct or incorrect decision given a set of similar past cases and how the advice of the AI agent turned out.

Three methods were provided as examples and evaluated based on nine combinations of AI techniques and benchmark data sets, as well as a single real-world decision-support use case. The results showed that case-based reasoning is a valid approach to accurately compute an AI agent's confidence. The use of synthetic data sets showed that the same approach is also accurate when the data distribution drifts over time or with sudden changes to the AI agent's behaviour (e.g., due to an update). In addition, these tests showed a higher degree of predictability than other common ways to compute confidences such as scoring [186], scaling [187], and voting [189] mechanisms.

Two human studies were performed to evaluate how easy to understand this case-based reasoning approach was (i.e., the method's interpretability). One study involved a controlled experiment and follow-up interview with domain experts. It showed that case-based reasoning matched their subjective intuition about how confidence should be computed. As such, they preferred this method over others. There was consensus that a case-based confidence estimation would make them more receptive to the use of an AI agent in their work and accept the support it offers. Finally, they viewed the possible explanations as valuable, particularly because they reminded them of the underlying case-based reasoning approach and helped them to assess the reliability of the confidence estimation.

The second study involved an online experiment with laypeople. The findings further indicated that, again, case-based reasoning matched their subjective intuition about confidence. Participants preferred a confidence explanation addressing past and current situations, as opposed to explanations addressing expected future elements that might influence the decision's outcome. Combined with those of the first study, these results indicate that a case-based reasoning approach to confidence estimation is transparent due to its underlying intuitive nature. It also offers explanations deemed valuable, as it refers to the past situations involved in the computation.

Chapter 6 addressed how an explanation could support a human contesting and altering an AI agent's decision, in particular when that decision has been deemed unfavourable or otherwise unreasonable. We focused particularly on cases in which the human has little to no direct influence on the AI agent itself and is subjected to its decision (e.g., when someone's loan or job application is rejected by an AI agent). This chapter formally defined the class of actionable explanations in such a context. Their purpose is to support humans taking effective action to alter an AI agent's decision to become more favourable or, otherwise, to effectively contest this decision through an accurate complaint to regulators.

This formal definition aimed to remove current ambiguity in what an actionable explanation is, with the purpose to provide direction. The formalisation should remove any ambiguity in concepts and requirements, something that is often the case within the field of XAI [32]. Such a formalisation allows researchers to assess if their method indeed generates actionable explanations through measurements addressing parts of the definition. This can further improve the much-needed measurements to assess progress in research towards actionable explanations, as it allows for the comparison between methods [33]. In turn, this supports a need for a more structured research community with a common research agenda [418, 419]. Finally, formal definitions of explanations open the possibility for AI agents to reason about their own explanations in symbolic and formal representation, enabling adaptive explanations [236].

To support these aims, the definition of an actionable explanation was divided in several properties. When an explanation would adhere to all properties, we deemed it actionable. This separation in distinct properties supports the identification of specific measurements, compare progress and methods, and the establishment of a research agenda. First, we defined that the explanation should be faithful to the AI agent's decision making and interpretable by the human. Second, the explanation should communicate alternative situations in which a different decision would be made. Such an alternative situation should be accompanied with suggestions on actions to realize those situations. Third, these actions should be feasible for that human, and the eventual alternative decisions should be deemed more favourable by them. These properties imply an accurate explanation and its communication, explicit communication to support actionability and a degree of personalization.

A literature review was conducted on methods that aim to generate actionable explanations. This review differed from others such as the one from Arrieta et. al [20] and Doshi-Velez et. al [33]. Ours aimed to provide an overview to which properties current methods adhere to and which are still open research challenges, this opposed to such other reviews that offer taxonomies or review the evaluation of methods. The review indicated an overall lack of provision of explicit action suggestions and personalization of such explanations. Currently, the literature on XAI methods is almost entirely limited to the faithful generation of these explanations, in particular the identification of alternative situations. However, the actual interpretability of these generated explanations is a matter of growing concern.

The topic of 'Contestable AI' is of growing interest to researchers [420], especially in healthcare [421]. A recent design-framework for instance, sets explanations as a requirement to enable human control and interventions [422]. Our work on actionable explanations offers a formalised framework to discuss and develop such explanations further.

To conclude, this second part identified and defined the novel explanation class of confidence explanations and defined and formalized the class of actionable explanations. Both aim to support humans to determine when and how to act appropriately when collaborating with an AI agent. With confidence explanations we enable humans to determine when it is best to rely on an AI agent's decision. With actionable explanations we support humans subjected to an AI agent's decision to effectively contest and alter an AI agent's decision. Both target an explicit information need needed that contribute to the responsible application of AI agents.

9.3 Findings: Part III

Explanations are a part of the human–AI collaboration that can take various forms depending on the use case. Thus, it is important to address the role of explanations in these various collaboration forms. In Part III, we address the following research question:

RQ 3: What is the role of explanations when human and AI agents collaborate on morally sensitive tasks?

In Chapter 7, we first extended a design method to human–AI collaboration, namely that of team design patterns [328]. Such patterns describe forms of collaboration by assigning certain work to humans, AI agents or both, depending on the situation. We reviewed this design approach for tasks involving morally sensitive decisions, and we extended the method's ontological decomposition of work with the notion of explaining moral context.

Four forms of collaboration were proposed that include the idea of explaining moral context. In the first form the human made all moral decision minimally supported by the AI agent performing supportive tasks such as collecting information, explanations were limited to disclose this collected information. The second form allowed the AI agent to support the human directly in their moral decisions by explaining relevant moral context and providing advice. The third form introduced the idea of the AI agent making all decisions but those for which it identified were beyond its capabilities, the role of explanations was to facilitate the handover of the decision by explaining why it lies outside of the AI agent's capabilities as well as the (morally) relevant context to make the decision. Finally, the fourth form introduced an AI agent that autonomously makes all decisions based on the elicited moral values from the human beforehand. The role of explanations in this collaboration form explained every decision made by the AI agent so the supervising human could adjust the agent in a next value elicitation process.

These four collaboration forms vary in the autonomy given to an AI agent and how the role of explanations varies with this degree of autonomy. We also addressed the benefit explanations can offer to establish such various forms in a high-risk application where morally sensitive decisions are involved. We especially addressed the benefit explanations about moral context can help humans make better moral decisions by preventing that humans overlook some vital aspects.

Our method to utilize team design patterns to address the role of explanations in human–AI collaborations enables a visual and comprehensible way to report on all effects of explanations within the socio-technical system formed by human, AI agent, task, and context. With this method we follow ideas similar to other design methods such the Socio-Cognitive Engineering method from Neerincx et. al [423, 77]. With the four proposed human–AI collaboration forms we discuss how explanation effects can be incorporated in a responsible design process of an AI agent where effects are well documented and evaluated.

Chapter 8 reports on a qualitative user study in which experts could experience three forms of human–AI collaboration in a medical triage task. These three were the following forms proposed in the previous chapter: 1) the AI agent as decision support to the human, 2) the AI agent acting autonomously but handing over decisions to the human when necessary, and 3) the AI agent acting autonomously under human supervision. These forms offered various degrees of human control over the AI agent by following a static (first and third form) or dynamic (second form) allocation of morally sensitive decisions. Each form used several classes of explanations: feature attributions, confidence explanations, and contrastive explanations. These explanations were applied to facilitate the human–AI collaboration during specific interactions. For example, a confidence explanation when the human proposed a certain decision to the AI agent.

The performed user study included first responders experiencing each collaboration form and the associated explanations. The experts reported that the simulated task was a valid representation of medical triage and that they felt most in control if that control immediately influenced what decisions the AI agent would make. In other words, they favoured the AI agent in a decision support role or to have the ability to take over decisions when deemed necessarily. They did not value their supervisory role with only the option to exercise control before and after the AI agent made all its decisions. The experts also did not feel any responsibility for the AI agent in their supervisory role.

Regarding the explanations, experts experienced them as valuable, though only when they felt they had the time to review them. In those cases, they mentioned that the information helped them determine when the AI agent could be trusted, and which tasks were best to take on themselves. When they were put under time pressure, they mostly emphasized the modality of the explanation (i.e., highly textual). They felt that reading and interpreting these explanations was time consuming and not worthwhile, as compared to making timely decisions and observing the AI agent's behaviour instead of reading about it in explanations. Finally, they also viewed parts of the interface that provided statistical data as explanation, although they were not intended as such.

These findings show the importance, possibilities, and difficulty of embedding explanations from an AI agent in an actual human–AI collaboration. Not only does the class and form of an explanation become important, but also its modality and embedding in the interface and task itself. It also points towards explanations that are adaptive to the context, for instance that presented information in an explanation is reduced when the human is under time pressure. This aligns with previous findings from the field of adaptive interface design that proposes the use of adaptive interfaces to match the cognitive and contextual demands humans experience during tasks [424].

To conclude, in this part we showed how to design a human–AI collaboration that accounts for a meaningful role for explanations. An evaluation of several of those designs illustrates the challenge of embedding explanations effectively in an interface, such that they do not become a hindrance instead of an aid. However, this evaluation has also revealed that humans perceive explanations as valuable for their collaboration with an AI agent, in terms of calibrating their trust and understanding the AI agent itself.

9.4 Limitations

The above reported results are limited in four aspects.

First, most of our performed quantitative user studies included proxies of real tasks, simulated AI agents and under-representative population samples. This was decided to guarantee a controlled experimental environment, to achieve a sufficient sample size and to be able to generalize outside of a particular use case. Although these limitations are common in user studies within the field of XAI, their results are limited in advancing the application of XAI in specific use cases [33]. These so-called human-grounded studies offer meaningful insight but do not explicitly contribute to the design choices we face in specific applications of AI agents. This opposed to application-grounded studies that involve the rich complexities of a realistic application of an AI agent. For instance, our evaluation on contrastive explanations from Chapter 2 indicates how such explanation can create for a more persuasive AI agent. However, further studies are required to what the extent this finding holds and under what kind of context such as the involved humans (e.g., domain experts, laypeople), task (e.g., low-risk, high-risk) and domain (e.g., diagnostic support, product recommendations). Although in some cases we did perform a study with domain experts such as in Chapter 5 and Chapter 8, these studies took the form of pilot studies and were of a more qualitative and exploratory nature.

Second, all explanations provided are one-shot and post-hoc explanations. One-shot refers to that limited interaction of the explanations and the lack of adaptation to that what was explained in the past. An understanding of a complex system is rarely obtained through a single explanation. Interaction between the AI agent and the human is needed to achieve a true understanding of the AI agent's functioning, especially if we regard a more long-term collaboration between human and AI agent [74]. Post-hoc explanations refer to those explanations that address a single decision after the AI agent arrived at it. There is growing concern that such an approach adds little to the trustworthiness of AI agents and the human ability to apply them responsibly [425, 426]. An increasing amount of merit is assigned to explanations that address how an AI agent is designed, developed, evaluated, and monitored. Such explanations need to be combined with post-hoc explanations, and ideally not independent of each other, to achieve trustworthy AI agents [24]. This thesis thus only addressed one potential topic of explanations, that of an AI agent's isolated decision.

Third, this thesis addressed the explanation's class (i.e., what is explained) and generative methods (i.e., how to acquire the explanation), we did not address the explanation's modality (i.e., how was explained). Insights from the field of human-computer interaction (HCI) were only integrated minimally in this thesis. Our studies however, showed the importance of an explanation's modality. For instance, the results from Chapter 7 and Chapter 8 indicate how explanations can be ignored or misused when not appropriately designed. The field of HCI offers expansive insights into the design of explanations [427]. In the past few years, the field's attention towards explanations pro-actively generated by AI agents has increased [428, 35]. Examples of related research are; the value of visual and textual contrastive explanations [429], the visualization of an AI agent's behaviour policy [430], the role of interaction design in AI contestability [431], and the co-creation of explanations in human-AI collaborations [432].

Fourth, the methods proposed to generate explanations in this thesis are all model agnostic. Although this allows such methods to be applied to a wide range of AI agents, they also suffer from a critical weakness [58]. Such explanations are mere approximations of an AI agent's decision-making process. Although such explanations should ideally be faithful to this process, meaning that they accurately approximate the process, they will never explain how the AI agent is making its decision. For example, we can approximate a neural network's decision boundary with a linear decision rule fairly accurate, but the decision rule will never capture the non-linear nuance of the decision boundary. Furthermore, as the AI agent grows so complex that no human can truly grasp its functioning, it becomes impossible to address how faithful an approximated explanation from a model-agnostic method is. After all, how can we determine that the approximation is faithful to the AI agent if we do not know what the explanation should be? Although this may not necessarily be an issue if it improves the human–AI collaboration, the problem remains that in most cases we still desire explanations to be empirically validated on their faithfulness lest they be perceived as false explanations.

9.5 Societal implications

The European Union recently published its approach to AI [433]. This draft for regulations dictates that AI agents should be safe and trustworthy. Their application should improve our lives without violating our fundamental rights. These proposed regulations underline both the promise and the risk of AI research. Rightly so, as we continue to experience the issues with the AI agents we apply increasingly.

A recent and relevant example is the use of a classification model used by the Dutch Tax and Customs Administration that discriminated against vulnerable minorities through ethnic profiling [1]. Over fifteen years, this model contributed to the accusation of around 27,000 parents of fraud, resulting in financial hardship for many. This scandal illustrates how AI agents can majorly impact people's lives.

This example and others contribute to our worries about using AI agents we do not understand. Such a lack of understanding prevents us from knowing if, when and how to use an AI agent responsibly and effectively. Luckily, this worry is shared among the research community, companies, and governments. As we performed this research, we discussed the topic of XAI with municipalities, governmental bodies, businesses (e.g., banks and software companies) and start-ups offering XAI solutions to such organisations. All believe that XAI can provide much-needed understanding of our AI agents. Consequently, we have experienced an increasing trend in the willingness to adopt explanation generating methods.

Considering the societal issues revolving around AI and the trend to adopt XAI methods, we feel responsible to reflect on our research and the societal implications of our findings. We want to reflect on our finding that explanations have both positive and negative effects, as determined by context. For instance, in Chapters 2 and 8 we discussed that many explanations we can currently generate only offer a limited understanding for non-AI experts. However, they can still significantly contribute to the calibration of trust and reliance on AI agents nonetheless, and offer various other benefits to a human–AI collaboration.

We believe that this finding is overlooked by the technology-centred research community of XAI. Many XAI methods are motivated solely by the authors' intuition, as we experienced in our own literature search (e.g., see Ehsan et al. [71] and Miller et al. [59]) and discussed by others such as Nauta et al. [41]). There is a clear societal risk to this perspective. It seems that within the field of XAI, the focus is on the question, "What *can* be explained?" instead of the question, "What *should* be explained?" Our findings indicate that the latter is equally important, if not more, if we want to ensure responsible use of AI agents. Without becoming more aware of all the effects explanations have, positive and negative, we risk issues like to those encountered in the Dutch childcare benefits scandal — worse, if the explanations inadvertently remove our current critical attitude towards AI agents.

Another finding we wish to reflect upon is that regarding the role of explanations. Currently, explanations are seen as a means to create trustworthy AI agents. However, explanations can be much more tailored to specific and concrete purposes, given the use case and joint task. Examples include the confidence explanations from Chapter 5 and actionable explanations from Chapter 6. In Chapters 7 and 8 we proposed a design method that embeds explanations in the human–AI collaboration with such specific purposes in mind. Our research illustrates that we can let AI agents generate explanations that are much more tailored, from explicitly calibrating human trust to help humans remain autonomy over their lives. This makes the field of XAI not only a solution to our worries about AI, but also an opportunity to expand on the possibilities of AI for our society.

The research should next regard how to thoroughly combine the technology-centred and human-centred perspectives in XAI research and their associated communities. Only then can we ensure that we responsibly apply the developed explanation generating methods. Luckily, this merging of communities is already occurring. To us personally this is most noticeable within the Netherlands Organisation for Applied Scientific Research (TNO), where our research was conducted. At TNO, the number of projects on XAI has increased. Their focus also moved from a few fundamental research projects developing new XAI methods to increasingly more applied research projects evaluating and validating such methods in particular use cases, especially in high-risk domains such as lifestyle advice [434], job vacancies [435], legal decision-support [436], human-machine teaming for the military [437], auditing the use of AI agents in the government [438], and meaningful and supervisory control over autonomous vessels [439].

Outside of TNO, we see various initiatives focusing on a more human-centred approach to XAI. Examples include the HI centre [440], the CEE-AI [441], the XAI Project [442]. Other initiatives aim to bolster Dutch and European research on AI with a human-centred perspective, often entailing some focus on XAI. These include the NLAIC [443], ELLIS Society [444], and CLAIRE [445].

We argue that these trends should continue and accelerate. Only through a combined human- and technology-centred perspective on XAI research will we learn what explanations an AI agent can and should convey to us for what purpose and in what context. We should not wait until the next scandal with an AI agent, as the next one might occur because of an irresponsible application of explanation generating methods. That such an error might occur out of ignorance is no excuse to those affected. Instead, we should directly involve companies and governments and help them to apply XAI research results responsibly in their human–AI collaboration.

CHAPTER 10

CONCLUSIONS

In this thesis, we aimed to design and develop explanations to support a responsible and effective collaboration between humans and AI agents. In this chapter we reflect on our main research question and provide an answer. We list our main conclusions for each part and end with a final conclusion. This conclusion summarizes the explanations we worked on, for what purpose they can be used given our results and how explanations can be incorporated in the design of human–AI collaborations.

We then discuss four main research challenges that we came across during our research; 1) perform more user studies, including studies grounded in realistic applications of AI agents, 2) base explanation generating methods more on fundamental mathematical theory to ensure their correctness, 3) incorporate a socio-technical system perspective in XAI research to identify and research more purposes of explanations, and 4) develop and propose concrete design methods and advice towards companies and governments on how XAI research results should be interpreted and applied in a responsible manner.

10.1 Main conclusions

Within this thesis we contributed to the following research question:

> What should an AI agent explain to humans,
> and how can it generate such explanations?

This question recognizes that an AI agent's explanation serves the humans who collaborate with it following the human-centred perspective on explainable AI (XAI) [236, 356]. Furthermore, the above question recognizes that this explanation should be feasible for the AI agent to generate, incorporating the technology-centred perspective on XAI. The two perspectives were combined in this thesis, and many of our insights relate to this combined perspective. We believe that our results show the merit of a combined perspective, where one accounts both for what is required and what is feasible.

An unequivocal definitive answer on the above research question is out of scope of a single thesis. The question describes the entire aim of the research field of XAI following a combined human- and technology-centred perspective. This thesis only contributed to this question within a specific scope defined by our research aim formulated as follows:

> To design and develop explanations that support a responsible
> and effective collaboration between humans and AI agents.

This scoped our research to the exploration of explanations offering value to the human–AI collaboration, in particular explanations that support a collaboration that leads to an effective and responsible completion of the joint task the human and AI agent face together. Within this scope a generic answer is best formulated as:

> An AI agent should provide contrastive, confidence and actionable explanations tailored to induce desirable effects in the humans that collaborate with that agent, where potential negative effects are either of limited consequence or mitigated in some way. Such explanations will vary widely given the needs posed by the role and background of the human and the shared context of both humans

and AI agent. Only through rigorous evaluations do we come to know what explanations are suited for what context. In turn, only through responsible design methods, such as the use of design patterns, can we incorporate those results in applied explainable AI agents.

Currently, explanations can be generated by approximating the AI agent's functioning with an accessible and interpretable model that can be used to base an explanation on. This approach is both efficient and sufficient, as humans rarely need to understand all aspects of an AI agent's functioning. Furthermore, such an approach is agnostic and future-proof due to being independent on technologies used to develop AI agents.

This is a generic answer that is not specific for a certain explanation. Instead, below we provide more detailed answers given the research questions and explanations addressed in each.

In Part I, we studied the effects of contrastive explanations, i.e., when an AI agent explains its reasons for arriving at a decision instead of another. Furthermore, we proposed methods how such explanations could be generated by an AI agent. One of our studies showed that a rule-based contrastive explanations support a human's understanding in why an AI agent made a certain decision. Furthermore, with several benchmark tests, we showed that rule-based contrastive explanations can be generated faithfully and efficiently independent from the internal functioning of the AI agent offering decision support. For AI agents that devise a plan of multiple actions over time, a pilot study showed that humans prefer an explanation that contrasts a proposed plan with an alternative one in terms of the difference in expected consequences and outcomes. We presented a method how such explanations could be generated by making use of the AI agent's model of the world combined with human expert knowledge to translate abstract and mathematical concepts into more human-readable concepts.

In Part II, we introduced confidence explanations, i.e., explanations that enable humans to determine whether the AI agent's decision can be trusted. In two studies we showed that human confidence in decisions can be interpreted as the likelihood of a decision being correct given past similar situations and the decisions made in those cases. We defined confidence explanations based on case-based reasoning to follow this interpretation. Not only did our benchmark tests show this approach to be an accurate estimate of the likelihood that a decision will be correct, but that this approach is also robust and predictable even under concept drift. Finally, we showed how case-based reasoning readily support human-interpretable textual explanations of the computed likelihood serving as a confidence estimate that was preferred by participants in our studies over others.

Aside from confidence explanations, in Part II we also defined the class of actionable explanations, i.e., explanations that explain how a human can contest an AI agent's decision by supporting them in taking the right action to alter this decision into a more favourable one. Six distinct properties were formally defined divided in three increasingly complex levels, based on a formal framework of contestability situated in a socio-technical system perspective. We stated that for an explanation to be deemed actionable – opposed to only offering epistemic value – it should not only be faithful to the AI agent's functioning and interpretable by the human, but also propose alternative situations in which the AI agent would make a more favourable decision. Furthermore, the explanation should disclose the actions needed to achieve those alternative situations. These should be actions that are feasible for the human to perform. With the help of a literature review, we concluded that the main future challenges to achieving actionable explanations are to address how to personalize the explanation to what decisions a human favours and what actions it can perform. Finally, we identified there is still relatively little research on how explanations should be communicated to be interpretable to humans.

Finally, in Part III, we investigated how explanations contribute to human–AI collaboration in morally sensitive tasks. We extended a design method for such collaboration forms based on the notion of design patterns to include the role and purpose of explanations in a morally sensitive context. Four of such forms were proposed, including the potential role of explanations in each. These forms varied the allocation of (morally sensitive) tasks between human and AI agent and how explanations can be used to support the handover of tasks as well as the human awareness on what the AI agent is doing. Three of these proposed collaboration forms, described as prototypical design patterns, were evaluated in a qualitative study with healthcare professionals and the morally sensitive task of medical triage. This evaluation substantiated our claim that explanations can have a multitude of effects on the human–AI collaboration, where the use-case determines if these effects are beneficial or detrimental to the collaboration. We showed that explanations should be carefully designed for and embedded in the human–AI collaboration. Otherwise, we risk that the use of explanations might introduce more negative effects (i.e., increased mental load) then positive ones (i.e., better task performance).

To conclude, our contribution to the posed research can be summarized that contrastive explanations can offer understanding in an AI agent's functioning, confidence explanations can assist in trust-calibration, and actionable explanations might improve human ability to contest AI agents. However, these explanations might also induce other effects. The ones were encountered include the creation of a more persuasive AI agent, an induced false sense of understanding, an increased human mental load, and an increased time to a decision. Whether such effects are detrimental to the human–AI collaboration and subsequently to the use-case, should be determined in the design process of the explainable AI agent. This design can be supported by creating design patterns that offer a list of evaluated effects, that can be used to make a responsible decision whether some explanation should be used by the AI agent in that envisioned application.

10.2 Future research

This thesis is a mere step towards new insights and technologies. As it closes, we contemplate the future of explainable AI research and report on challenges left open. We review what must be done to be able to design and develop explanations that support a responsible and effective collaboration between humans and AI agents.

We are strong believers that explanations should be evaluated rigorously in human studies. This is due to the first challenge we wish to raise, that is that we still know little about the multitude of effects causes by an explanation about an AI agent's functioning. In our studies we found that explanations have a variety of effects, including effects that were unintended. We argue that the research community has much to gain from conducting more human studies, both in approximations and realistic applications of AI agents. Such studies would enable the development of theoretical models on how explanations and effects are linked given contextual aspects such as the task, human expertise, and the roles of the human and AI agent. These models are requested for to help structure the research on explainable AI agents [61, 32]. To enable to community to conduct more human studies we need to create and validate metrics that measure the effects of explanations. Furthermore, we call for the experimentation and development of study designs to evaluate explanations. It has been raised that the research community, due to its highly multi-disciplinary nature, has trouble designing and understanding human studies from which reliable conclusions can be drawn [59]. Furthermore, studying the effects of explanations on understanding brings a profound learning effect in studies and might even require long-term studies. Both adds to the difficulty of designing an appropriate study. We should look upon other fields experienced in conducting human studies and apply or adapt their practices to the field of XAI research.

The second research challenge we identify is to research more fundamental theories on how we can develop methods that can generate explanations about AI agents so complex that they cannot be readily understood by any human. Currently, such explanations are generated using the approach of surrogate models, that is, models that are more interpretable and accessible that are used to approximate the much more complex AI agent [140]. Within this thesis we developed several methods based on surrogate models, and as others, we evaluated their correctness with the help of benchmarks to measure their efficiency and faithfulness. However, such benchmarks do not guarantee that the method will work for every AI agent and its application. Although surrogate model methods allow an AI agent to explain itself independent of its own complexity, the fundamental issue with these methods is that we cannot determine when this necessarily simplified explanation correctly describes this AI agent that is too complex for us to understand as is. There is little to no attention within the field of XAI research to this fundamental issue. We call for the development of more fundamental theories, relying on mathematics and philosophy, how to generate sufficient descriptions that act as correct explanations of the much more complex system they explain.

The third research challenge relates to widen the scope of XAI research. Currently, an explainable AI agent is mostly seen to provide a necessary understanding to make it trustworthy. However, explanations can serve many more purposes in the human–AI collaboration. With a more socio-technical system perspective, where human and AI agent are viewed as two collaborating and interacting agents within a certain context [77, 356], the field can explore a wider variety of uses of explanations. In particular, how explanations from an AI agent can add to the functionalities of the AI agent. For instance, explanations might educate the human in a domain, help discover new (causal) relations, or improve their creative thinking and problem solving. Not only can this socio-technical system perspective lead to new uses of explanations, but it will also underline the necessity for research to focus on how explanations are communicated. There is little attention to the development of interface and interaction designs that communicate explanations in an interpretable way [35]. The research in this thesis had similar limitations, although our final pilot study indicated that the modality of explanations is pivotal in whether these explanations add value to the human–AI collaboration. Thus, if we want to achieve an effective and responsible human–AI collaboration with the help of explanations, more research should be conducted to include more purposes of explanations and how explanations should be communicated.

Finally, the fourth research challenge is the development of best practices and policies that dictate how to design, implement, validate and maintain explainable AI agents. Our conversations with companies and government institutes made us realize that their biggest question is how to utilize all the progress that is made within the field of XAI. They wonder how this should be incorporated in their decision-making when and where to apply AI agents, especially with respect to upcoming regulations and the current mixed societal view towards AI agents. If they decide to apply an AI agent or already do so, they wonder how to add explanations to it. They are uncertain about the design choices that need to be made and, when made, how they should be incorporated in the AI agent such that the explanations have a positive effect. The above research challenges all contribute to this need. Through explanations validated with user studies and generated through methods that are fundamentally sound, we can develop design methods when and where to apply which explanations and what methods that can generate them. With a widened scope of XAI research, we can furthermore expand the functionalities of AI agents so that, when applied, they bring more value to our society. However, to do so we must keep reflecting on our research progress with the need of companies and governments in mind. Given our experience, their current and greatest need is to have tools and methods available that help them understand and incorporate the progress of XAI research into their applications in a responsible manner.

For certain, plenty remains before we can truly speak of a self-explaining AI agent whose explanations contribute and enable an effective and responsible collaboration with humans. Thankfully, the research community is maturing, as methods are being developed, experiments are conducted, and different communities are increasingly involved. Rapid progress is being made towards the common goal of AI agents that can explain themselves.

ACKNOWLEDGEMENTS

Not only did I learn much about explainable AI, I also learned much about myself and doing research during the process of creating this thesis. Most of all, I learned that being a researcher is about interacting with others. Our explanations not only convey understanding to others, they create it in oneself as well. Many people contributed to this thesis, and without them this thesis would not exist in its current form.

At the end of this journey, I begin by thanking Jurriaan van Diggelen, Mark Neerincx and Catholijn Jonker. Your guidance and mentoring helped me become a researcher. Jurriaan, without you making room for my research at TNO, none of this would have happened. You always had confidence in me and allowed me to follow my own interests and ideas — even if they went nowhere at times. Mark, a long time ago you told me that combining applied research with a PhD is very much possible at TNO. This suggestion motivated me to start this journey. Throughout it, your feedback and ideas always made me look at the real implications of my work. Catholijn, your view on research encouraged me always to go a step further. You once told me that being a PhD candidate is about learning to become an independent researcher. That you happen to become an expert in a field is just the cherry on top. You might not realize it, but this statement always kept me motivated.

I am also grateful for the opportunity both TNO and the TuD offered me. My PhD journey was an odd one as a full-time researcher at TNO, and only with the warm hospitality of both TNO and TuD was I able to conduct my research. My gratitude to everyone in both organisations, from project leaders to managers and fellow PhD students to staff, for their open minds that allowed me to conduct research through much trial and error. In particular, I am grateful for the funding TNO supplied to do my own research.

Being part of a defence committee always seemed to me a great responsibility and still does. My gratitude goes towards all of my committee members. I know how busy one can become and how much time reviewing someone's work can cost. My deepest gratitude for the time you invested in me and my work.

I was lucky to be surrounded by great colleagues and researchers. To be able to do my research at TNO felt like my greatest advantage, although many raise it as a challenge. Many of my colleagues expressed how difficult it would be to do a PhD in our projects. However, I experienced the contrary as the people I had the privilege of working with added much to my research. My heartfelt thanks goes to all of those I worked with in countless projects. In particular Tjeerd, Tjalling, Sabine, Karel, Marieke, and Hans, I had the honour of writing many papers with you and learn from your unique perspectives. This all enriched my research more than I ever hoped for.

Similarly, I would like to thank the students I had the privilege to supervise. Thank you, Marcel, Elisabeth, Chantal, Manon, Soroosh, Wouter and Dhivin. I hope you learned just as much from me as I did from you.

My fellow PhD students at Delft showed me a joy I never expected when working in the academic world. Eli, Franziska, and Emma, your advice helped me face the practicalities of doing a PhD at Delft and glimpse how a full-time PhD at a university looks like. I would also like to thank Simon as a regular coach in some of the most interesting courses I took. They were always useful and great fun.

Much of my research is human-centred, and thus I spoke with many people about it as I went along. Without all the participants in my experiments and interviews with domain experts, I would not have figured how much of a double-sided coin Explainable AI really is. Similarly, without the discussions I had with people from companies and governmental bodies alike, I would not have fully understood the necessity of it nor its ramifications. If my research ends up helping just some of you only a little, I will consider it a great success.

In closing, I would like to thank everyone who supported me. Research can be hard, and I was lucky to have friends to support me. Alexander, even though we see each other seldom, our mail conversations continue to inspire; you are a present-day pen pal. Especially when I had just started my journey. Tjeerd, Tjalling, Sabine and Ruben, you are not just colleagues but friends. Our planned and unplanned "play dates" helped me unwind. At times, they changed into complaining sessions; however, those helped me to cope with all the stress. Finally, our countless whiteboard sessions helped me to understand my own thoughts, even though I suspect they just confused yours.

My gratitude to my friends Rick, Rianne, Pim and Manon. Our Dungeons & Dragons sessions and visits to zoos were a nice distraction from my work. Although you arrived late in my journey, Dejan and Sonja, our late-night game sessions prevented me from working too much. I also learned much about Serbian culture, hvala!

As so often, my family played an important role behind the scenes. I thank my parents for helping me become the person I am today. Mom, you always motivated me to aim high and to be ambitious which motivated me to get as far as I did. Dad, your perspective in life helped me become a decent and hardworking person, and I got my work ethics who helped me through my PhD from you. I cannot forget my big sister and your everlasting support, even though we live far apart. Our lives have not been easy at times, yours the least. Your support means all the more to me. I hope you realize how much I love you all.

Youetta, my love. You kept up with me all these years during this journey and all of its ups and downs. We got married, bought our first house and adopted countless of pets and plants. I know that I can complain much, and my work could make things more difficult at times; however, you were and are always there for me. Thank you, my love.

BIBLIOGRAPHY

[1] Amnesty International. *Amnesty report childcare benefits scandal*. 2021. URL: https://www.amnesty.nl/content/uploads/2021/10/NL_Xenophobe-Machines_Amnesty-International_EUR-35_4686_2021.pdf. (accessed: 18-8-2022).

[2] Randall Davis. 'Knowledge-based systems'. In: *Science* 231.4741 (1986), pp. 957–963.

[3] William R Swartout and Johanna D Moore. 'Explanation in second generation expert systems'. In: *Second generation expert systems*. Springer, 1993, pp. 543–585.

[4] Maaike Harbers. 'Explaining agent behavior in virtual training'. PhD thesis. University Utrecht, 2011.

[5] Richard W Southwick. 'Explaining reasoning: an overview of explanation in knowledge-based systems'. In: *The knowledge engineering review* 6.1 (1991), pp. 1–19.

[6] Melda Yucel, Gebrail Bekdaş and Sinan Melih Nigdeli. 'Review and Applications of Machine Learning and Artificial Intelligence in Engineering: Overview for Machine Learning and AI'. In: *Artificial Intelligence and Machine Learning Applications in Civil, Mechanical, and Industrial Engineering* (2020), pp. 1–12.

[7] David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf and Guang-Zhong Yang. 'XAI—Explainable artificial intelligence'. In: *Science Robotics* 4.37 (2019).

[8] Atoosa Kasirzadeh. 'Reasons, Values, Stakeholders: A Philosophical Framework for Explainable Artificial Intelligence'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 14–14.

[9] Yunhe Pan. 'Heading toward artificial intelligence 2.0'. In: *Engineering* 2.4 (2016), pp. 409–413.

[10] Emily Sullivan. 'Understanding from machine learning models'. In: *The British Journal for the Philosophy of Science* (2020).

[11] European Committee. *Data protection in the EU*. 2016. URL: https://ec.europa.eu/info/law/law-topic/data-protection/data-protection-eu_en. (accessed: 18-8-2022).

[12] European Committee. *Proposal for a Regulation laying down harmonised rules on artificial intelligence*. 2021. URL: https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence. (accessed: 18-8-2022).

[13] ProPublica. *Machine Bias*. 2016. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. (accessed: 18-8-2022).

[14] Forbes. *Google Photos Tags Two African-Americans As Gorillas Through Facial Recognition Software*. 2015. URL: https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/. (accessed: 18-8-2022).

[15] Reuters. *Amazon scraps secret AI recruiting tool that showed bias against women*. 2018. URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. (accessed: 18-8-2022).

[16] Krishna Gade, Sahin Geyik, Krishnaram Kenthapadi, Varun Mithal and Ankur Taly. 'Explainable AI in industry: practical challenges and lessons learned'. In: *Companion Proceedings of the Web Conference 2020*. 2020, pp. 303–304.

[17] M. T. Ribeiro, S. Singh and C. Guestrin. '"Why Should I Trust You?": Explaining the Predictions of Any Classifier'. In: *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144.

[18] Leilani H Gilpin, Cecilia Testart, Nathaniel Fruchter and Julius Adebayo. 'Explaining explanations to society'. In: *NeurIPS Workshop on Ethical, Social and Governance Issues in AI*. 2019.

[19] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro and Daniel Weld. 'Does the whole exceed its parts? the effect of ai explanations on complementary team performance'. In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–16.

[20] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins et al. 'Explainable Artificial

Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI'. In: *Information Fusion* 58 (2020), pp. 82–115.

[21] Mireia Ribera and Agata Lapedriza. 'Can we do better explanations? A proposal of user-centered explainable AI.' In: *IUI Workshops*. Vol. 2327. 2019, p. 38.

[22] Joachim de Greeff, Maaike HT de Boer, Fieke HJ Hillerström, Freek Bomhof, Wiard Jorritsma and Mark A Neerincx. 'The FATE System: FAir, Transparent and Explainable Decision Making.' In: *AAAI Spring Symposium: Combining Machine Learning with Knowledge Engineering*. 2021.

[23] Andrés Páez. 'The pragmatic turn in explainable artificial intelligence (XAI)'. In: *Minds and Machines* 29.3 (2019), pp. 441–459.

[24] Brent Mittelstadt, Chris Russell and Sandra Wachter. 'Explaining explanations in AI'. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 279–288.

[25] Sanjay Krishnan and Eugene Wu. 'Palm: Machine learning explanations for iterative debugging'. In: *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*. 2017, pp. 1–6.

[26] Todd Kulesza, Simone Stumpf, Weng-Keen Wong, Margaret M Burnett, Stephen Perona, Andrew Ko and Ian Oberst. 'Why-oriented end-user debugging of naive Bayes text classification'. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 1.1 (2011), pp. 1–31.

[27] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović et al. 'AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias'. In: *IBM Journal of Research and Development* 63.4/5 (2019), pp. 4–1.

[28] Christian Meske, Enrico Bunde, Johannes Schneider and Martin Gersch. 'Explainable artificial intelligence: objectives, stakeholders, and future research opportunities'. In: *Information Systems Management* (2021), pp. 1–11.

[29] Sandra Wachter, Brent Mittelstadt and Chris Russell. 'Counterfactual explanations without opening the black box: Automated decisions and the GDPR'. In: *Harv. JL & Tech.* 31 (2017), p. 841.

[30] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou and Murat Kantarcioglu. ‘Does Explainable Artificial Intelligence Improve Human Decision-Making?’ In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. 2021, pp. 6618–6626.

[31] Karel van den Bosch, Tjeerd Schoonderwoerd, Romy Blankendaal and Mark Neerincx. ‘Six challenges for human-AI Co-learning’. In: *International Conference on Human-Computer Interaction*. Springer. 2019, pp. 572–589.

[32] Zachary C Lipton. ‘The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery’. In: *Queue* 16.3 (2018), pp. 31–57.

[33] F Doshi-Velez and B Kim. ‘Towards A Rigorous Science of Interpretable Machine Learning’. In: *arXiv preprint arXiv:1702.08608* (2017). arXiv: 1702.08608.

[34] Wojciech Samek, Thomas Wiegand and Klaus-Robert Müller. ‘Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models’. In: *ITU Journal: ICT Discoveries* (2017).

[35] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim and Mohan Kankanhalli. ‘Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda’. In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. 2018, pp. 1–18.

[36] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal. ‘Explaining explanations: An overview of interpretability of machine learning’. In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 2018, pp. 80–89.

[37] B. Herman. ‘The Promise and Peril of Human Evaluation for Model Interpretability’. In: *Proceedings of the Symposium on Interpretable Machine Learning*. 2017.

[38] Amina Adadi and Mohammed Berrada. ‘Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI)’. In: *IEEE Access* 6 (2018), pp. 52138–52160.

[39] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti and Dino Pedreschi. ‘A survey of methods for explaining black box models’. In: *ACM computing surveys (CSUR)* 51.5 (2018), pp. 1–42.

[40] Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. 'Explainable ai: A review of machine learning interpretability methods'. In: *Entropy* 23.1 (2021), p. 18.

[41] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen and Christin Seifert. 'From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI'. In: *arXiv preprint arXiv:2201.08164* (2022).

[42] Tim Miller. 'Contrastive explanation: A structural-model approach'. In: *The Knowledge Engineering Review* 36 (2021).

[43] Georgios Papadopoulos, Peter J Edwards and Alan F Murray. 'Confidence estimation methods for neural networks: A practical comparison'. In: *IEEE Transactions on Neural Networks* 12.6 (2001), pp. 1278–1287.

[44] Umang Bhatt, Adrian Weller and José MF Moura. 'Evaluating and aggregating feature-based model explanations'. In: *Proceedings of the 29th International Joint Conference on Artificial Intelligence*. 2020.

[45] Carrie J Cai, Jonas Jongejan and Jess Holbrook. 'The effects of example-based explanations in a machine learning interface'. In: *Proceedings of the 24th international conference on intelligent user interfaces*. 2019, pp. 258–262.

[46] Cynthia Rudin and Yaron Shaposhnik. 'Globally-consistent rule-based summary-explanations for machine learning models: application to credit-risk evaluation'. In: *Available at SSRN 3395422* (2019).

[47] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper and Haiyi Zhu. 'Explaining decision-making algorithms through UI: Strategies to help non-expert stakeholders'. In: *Proceedings of the 2019 chi conference on human factors in computing systems*. 2019, pp. 1–12.

[48] Josua Krause, Adam Perer and Kenney Ng. 'Interacting with predictions: Visual inspection of black-box machine learning models'. In: *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2016, pp. 5686–5697.

[49] Kacper Sokol and Peter A Flach. 'Conversational Explanations of Machine Learning Predictions Through Class-contrastive Counterfactual Statements.' In: *IJCAI*. 2018, pp. 5785–5786.

[50] Richard Dazeley, Peter Vamplew, Cameron Foale, Charlotte Young, Sunil Aryal and Francisco Cruz. 'Levels of explainable artificial intelligence for human-aligned conversational explanations'. In: *Artificial Intelligence* 299 (2021), p. 103525.

[51] Mengnan Du, Ninghao Liu and Xia Hu. 'Techniques for interpretable machine learning'. In: *Communications of the ACM* 63.1 (2019), pp. 68–77.

[52] Leo Breiman, Jerome H Friedman, Richard A Olshen and Charles J Stone. *Classification and regression trees*. Routledge, 2017.

[53] Alberto Blanco-Justicia and Josep Domingo-Ferrer. 'Machine learning explainability through comprehensible decision trees'. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer. 2019, pp. 15–26.

[54] Javier Antoran, Umang Bhatt, Tameem Adel, Adrian Weller and José Miguel Hernández-Lobato. 'Getting a CLUE: A Method for Explaining Uncertainty Estimates'. In: *International Conference on Learning Representations*. 2020.

[55] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal and Su-In Lee. 'From local explanations to global understanding with explainable AI for trees'. In: *Nature machine intelligence* 2.1 (2020), pp. 56–67.

[56] Goutham Ramakrishnan, Yun Chan Lee and Aws Albarghouthi. 'Synthesizing action sequences for modifying model decisions'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 04. 2020, pp. 5462–5469.

[57] Fulton Wang and Cynthia Rudin. 'Falling rule lists'. In: *Artificial Intelligence and Statistics*. Vol. 38. 2015.

[58] Cynthia Rudin. 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead'. In: *Nature Machine Intelligence* 1.5 (2019), pp. 206–215.

[59] T. Miller, P. Howe and L. Sonenberg. 'Explainable AI: Beware of Inmates Running the Asylum'. In: *Proceedings of the International Joint Conference Artificial Intelligence (IJCAI)*. 2017, pp. 36–41.

[60] Douglas Cirqueira, Dietmar Nedbal, Markus Helfert and Marija Bezbradica. 'Scenario-based requirements elicitation for user-centric explainable ai'. In: *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer. 2020, pp. 321–341.

[61] Robert R Hoffman, Shane T Mueller, Gary Klein and Jordan Litman. ‘Metrics for explainable AI: Challenges and prospects’. In: *arXiv preprint arXiv:1812.04608* (2018).

[62] Maximilian A Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith and Dimitri Bohlender. ‘Explainability as a non-functional requirement’. In: *2019 IEEE 27th International Requirements Engineering Conference (RE)*. IEEE. 2019, pp. 363–368.

[63] Q. V. Liao, D. Gruen and S. Miller. ‘Questioning the AI: informing design practices for explainable AI user experiences’. In: *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 2020, pp. 1–15.

[64] Larissa Chazette and Kurt Schneider. ‘Explainability as a non-functional requirement: challenges and recommendations’. In: *Requirements Engineering* 25.4 (2020), pp. 493–514.

[65] Sahil Verma, Aditya Lahiri, John P Dickerson and Su-In Lee. ‘Pitfalls of Explainable ML: An Industry Perspective’. In: *Proceedings of the Fifth Conference on Machine Learning and Systems*. 2021.

[66] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger and Andreas Butz. ‘I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI’. In: *26th International Conference on Intelligent User Interfaces*. 2021, pp. 307–317.

[67] Tomáš Kliegr, Štěpán Bahník and Johannes Fürnkranz. ‘A review of possible effects of cognitive biases on interpretation of rule-based machine learning models’. In: *Artificial Intelligence* (2021), p. 103458.

[68] Tim Miller. ‘Explanation in artificial intelligence: Insights from the social sciences’. In: *Artificial intelligence* 267 (2019), pp. 1–38.

[69] Bertram F Malle. *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Mit Press, 2006.

[70] Matthew Johnson and Alonso Vera. ‘No AI is an island: the case for teaming intelligence’. In: *AI Magazine* 40.1 (2019), pp. 16–28.

[71] Upol Ehsan and Mark O Riedl. ‘Human-centered explainable ai: Towards a reflective sociotechnical approach’. In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 449–466.

[72] Maartje MA De Graaf and Bertram F Malle. ‘How people explain action (and autonomous intelligent systems should too)’. In: *2017 AAAI Fall Symposium Series*. 2017.

[73] Sule Anjomshoae, Amro Najjar, Davide Calvaresi and Kary Främling. 'Explainable agents and robots: Results from a systematic literature review'. In: *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2019, pp. 1078–1088.

[74] Ewart J de Visser, Marieke MM Peeters, Malte F Jung, Spencer Kohn, Tyler H Shaw, Richard Pak and Mark A Neerincx. 'Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams'. In: *International journal of social robotics* 12.2 (2020), pp. 459–478.

[75] Ivan Contreras and Josep Vehi. 'Artificial intelligence for diabetes management and decision support: literature review'. In: *Journal of medical Internet research* 20.5 (2018), e10775.

[76] Ioannis Kavakiotis, Olga Tsave, Athanasios Salifoglou, Nicos Maglaveras, Ioannis Vlahavas and Ioanna Chouvarda. 'Machine learning and data mining methods in diabetes research'. In: *Computational and structural biotechnology journal* 15 (2017), pp. 104–116.

[77] Mark A Neerincx, Willeke van Vught, Olivier Blanson Henkemans, Elettra Oleari, Joost Broekens, Rifca Peters, Frank Kaptein, Yiannis Demiris, Bernd Kiefer, Diego Fumagalli et al. 'Socio-Cognitive Engineering of a Robotic Partner for Child's Diabetes Self-Management'. In: *Frontiers in Robotics and AI* 6 (2019).

[78] Bradley Hayes and Julie A Shah. 'Improving robot controller transparency through autonomous policy explanation'. In: *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE. 2017, pp. 303–312.

[79] Tathagata Chakraborti, Anagha Kulkarni, Sarath Sreedharan, David E Smith and Subbarao Kambhampati. 'Explicability? legibility? predictability? transparency? privacy? security? the emerging landscape of interpretable agent behavior'. In: *Proceedings of the International Conference on Automated Planning and Scheduling*. 29. 2019, pp. 86–96.

[80] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber and Katelyn Procci. 'Intelligent agent transparency in human–agent teaming for Multi-UxV management'. In: *Human factors* 58.3 (2016), pp. 401–415.

[81] Kristen Stubbs, Pamela J Hinds and David Wettergreen. 'Autonomy and common ground in human-robot interaction: A field study'. In: *IEEE Intelligent Systems* 22.2 (2007), pp. 42–50.

[82] Todd Kulesza, Margaret Burnett, Weng-Keen Wong and Simone Stumpf. 'Principles of explanatory debugging to personalize interactive machine learning'. In: *Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM. 2015, pp. 126–137.

[83] Or Biran and Courtenay Cotton. 'Explanation and justification in machine learning: A survey'. In: *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. 2017, p. 1.

[84] Mustafa Bilgic and Raymond J Mooney. 'Explaining recommendations: Satisfaction vs. promotion'. In: *Beyond Personalization Workshop, IUI*. Vol. 5. 2005, p. 153.

[85] Upol Ehsan, Brent Harrison, Larry Chan and Mark O Riedl. 'Rationalization: A neural machine translation approach to generating natural language explanations'. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM. 2018, pp. 81–87.

[86] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele and Trevor Darrell. 'Generating visual explanations'. In: *European Conference on Computer Vision*. Springer. 2016, pp. 3–19. ISBN: 9783319464923. DOI: 10.1007/978-3-319-46493-0_1. eprint: 1603.08507.

[87] Jonathan L Herlocker, Joseph A Konstan and John Riedl. 'Explaining collaborative filtering recommendations'. In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. ACM. 2000, pp. 241–250.

[88] L Richard Ye and Paul E Johnson. 'The impact of explanation facilities on user acceptance of expert systems advice'. In: *Mis Quarterly* (1995), pp. 157–172.

[89] Jianlong Zhou, Zhidong Li, Huaiwen Hu, Kun Yu, Fang Chen, Zelin Li and Yang Wang. 'Effects of influence on user trust in predictive decision making'. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 2019, pp. 1–6.

[90] Shlomo Berkovsky, Ronnie Taib and Dan Conway. 'How to recommend?: User trust factors in movie recommender systems'. In: *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. ACM. 2017, pp. 287–300.

[91] Daniel Holliday, Stephanie Wilson and Simone Stumpf. 'User trust in intelligent systems: A journey over time'. In: *Proceedings of the 21st International Conference on Intelligent User Interfaces*. ACM. 2016, pp. 164–168.

[92] Florian Nothdurft, Tobias Heinroth and Wolfgang Minker. 'The impact of explanation dialogues on human-computer trust'. In: *International Conference on Human-Computer Interaction*. Springer. 2013, pp. 59–67.

[93] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman and Finale Doshi-Velez. 'An evaluation of the human-interpretability of explanation'. In: *Proceedings of the Conference on Neural Information Processing System*. 2019.

[94] M Joppe. *The Research Process. Retrieved February 25, 1998*. 2000.

[95] Ellen A Drost et al. 'Validity and reliability in social science research'. In: *Education Research and perspectives* 38.1 (2011), p. 105.

[96] Jerome Kirk, Marc L Miller and Marc Louis Miller. *Reliability and validity in qualitative research*. Vol. 1. Sage, 1986.

[97] Peter Lipton. 'Contrastive explanation'. In: *Royal Institute of Philosophy Supplements* 27 (1990), pp. 247–266.

[98] Brian Y Lim and Anind K Dey. 'Assessing demand for intelligibility in context-aware applications'. In: *Proceedings of the 11th international conference on Ubiquitous computing*. ACM. 2009, pp. 195–204.

[99] L Karl Branting. 'Building explanations from rules and structured cases'. In: *International journal of man-machine studies* 34.6 (1991), pp. 797–837.

[100] J van der Waa, M Robeer, J van Diggelen, M Brinkhuis and M Neerincx. 'Contrastive Explanations with Local Foil Trees'. In: *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*. Vol. 37. 2018.

[101] Nahla Barakat and Joachim Diederich. 'Eclectic rule-extraction from support vector machines'. In: *International Journal of Computational Intelligence* 2.1 (2005), pp. 59–62.

[102] Allen Newell, Herbert Alexander Simon et al. *Human problem solving*. Vol. 104. 9. Prentice-hall Englewood Cliffs, NJ, 1972.

[103] Michelene TH Chi, Miriam Bassok, Matthew W Lewis, Peter Reimann and Robert Glaser. 'Self-explanations: How students study and use examples in learning to solve problems'. In: *Cognitive science* 13.2 (1989), pp. 145–182.

[104] Alexander Renkl. 'Worked-out examples: Instructional explanations support learning by self-explanations'. In: *Learning and instruction* 12.5 (2002), pp. 529–556.

[105] Irit Peled and Orit Zaslavsky. ‘Counter-Examples That (Only) Prove and Counter-Examples That (Also) Explain.’ In: *FOCUS on Learning Problems in mathematics* 19.3 (1997), pp. 49–61.

[106] Ajaya Adhikari, David MJ Tax, Riccardo Satta and Matthias Faeth. ‘LEAF-AGE: Example-based and Feature importance-based Explanations for Black-box ML models’. In: *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE. 2019, pp. 1–7.

[107] Jacob Bien, Robert Tibshirani et al. ‘Prototype selection for interpretable classification’. In: *The Annals of Applied Statistics* 5.4 (2011), pp. 2403–2424.

[108] Been Kim, Cynthia Rudin and Julie A Shah. ‘The bayesian case model: A generative approach for case-based reasoning and prototype classification’. In: *Advances in Neural Information Processing Systems*. 2014, pp. 1952–1960.

[109] Been Kim, Rajiv Khanna and Oluwasanmi O Koyejo. ‘Examples are not enough, learn to criticize! criticism for interpretability’. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2280–2288.

[110] Robert K Atkinson. ‘Optimizing learning from examples using animated pedagogical agents.’ In: *Journal of Educational Psychology* 94.2 (2002), p. 416.

[111] Michael J Pazzani. ‘Representation of electronic mail filtering profiles: a user study’. In: *Proceedings of the 5th international conference on Intelligent user interfaces*. 2000, pp. 202–206.

[112] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan and Jonathan Herlocker. ‘Interacting meaningfully with machine learning systems: Three experiments’. In: *International Journal of Human-Computer Studies* 67.8 (2009), pp. 639–662.

[113] Adrian Bussone, Simone Stumpf and Dympna O’Sullivan. ‘The role of explanations on trust and reliance in clinical decision support systems’. In: *2015 International Conference on Healthcare Informatics*. IEEE. 2015, pp. 160–169.

[114] Brian Y Lim, Anind K Dey and Daniel Avrahami. ‘Why and why not explanations improve the intelligibility of context-aware intelligent systems’. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2009, pp. 2119–2128.

[115] Judea Pearl et al. ‘Causal inference in statistics: An overview’. In: *Statistics surveys* 3 (2009), pp. 96–146.

[116] Vibhu O Mittal and Cécile L Paris. 'Generating explanations in context: The system perspective'. In: *Expert Systems with Applications* 8.4 (1995), pp. 491–503.

[117] Nancy J Cooke, Steven M Shope, LRE Schiflett, E Salas and MD Coovert. 'Designing a synthetic task environment'. In: *Scaled worlds: Development, validation, and application* (2004), pp. 263–278.

[118] Valentine U Odili, Paul D Isiboge and Aihanuwa Eregie. 'Patients' knowledge of diabetes mellitus in a Nigerian city'. In: *Tropical Journal of Pharmaceutical Research* 10.5 (2011), pp. 637–642.

[119] Gabriele Paolacci, Jesse Chandler and Panagiotis G Ipeirotis. 'Running experiments on amazon mechanical turk'. In: *Judgment and Decision making* 5.5 (2010), pp. 411–419.

[120] Andrea Papenmeier, Gwenn Englebienne and Christin Seifert. 'How model accuracy and explanation fidelity influence user trust'. In: *arXiv preprint arXiv:1907.12652* (2019).

[121] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh and Himabindu Lakkaraju. 'How can we fool LIME and SHAP? Adversarial Attacks on Post hoc Explanation Methods'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2019.

[122] Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan and Madeleine Udell. '"Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations'. In: *Proceedings of the International Conference on Machine Learning AI for Social Good Workshop*. 2019.

[123] V Buch, G Varughese and M Maruthappu. 'Artificial intelligence in diabetes care'. In: *Diabetic Medicine* 35.4 (2018), pp. 495–497.

[124] Monika Reddy, Sian Rilstone, Philippa Cooper and Nick S Oliver. 'Type 1 diabetes in adults: supporting self management'. In: *Bmj* 352 (2016), p. i998.

[125] I. Bosch. *Het Groot Diabetesboek*. Mension B.V., 2013.

[126] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal and Heimo Müller. 'Causability and explainability of artificial intelligence in medicine'. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9.4 (2019), e1312.

[127] RC Bogdan and SK Biklen. 'Qualitative research in (validation) and qualitative (inquiry) studies'. In: *It is a method-appropriate education: An introduction to theory and methods* (2006).

[128] Carl J Huberty and John D Morris. ‘Multivariate analysis versus multiple univariate analyses.’ In: (1992).

[129] Geoffrey Keppel. *Design and analysis: A researcher’s handbook*. Prentice-Hall, Inc, 1991.

[130] Esra Yigit and Fikri Gokpinar. ‘A simulation study on tests for one-way ANOVA under the unequal variance assumption’. In: *Commun Fac Sci Univ Ankara, Ser A* 1 (2010), pp. 15–34.

[131] K. Crawford. ‘Artificial Intelligence’s White Guy Problem’. In: *The New York Times* (2016).

[132] A. Caliskan, J. J. Bryson and A. Narayanan. ‘Semantics Derived Automatically from Language Corpora Contain Human-Like Biases’. In: *Science* 356.6334 (2017), pp. 183–186.

[133] S. Barocas and A. D. Selbst. ‘Big Data’s Disparate Impact’. In: *Cal. L. Rev.* 104 (2016), p. 671.

[134] C. Coglianese and D. Lehr. ‘Regulating by Robot: Administrative Decision Making in the Machine-Learning Era’. In: *Geo. LJ* 105 (2016), p. 1147.

[135] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton and Derek Roth. ‘A comparative study of fairness-enhancing interventions in machine learning’. In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 329–338.

[136] T. Kulesza, S. Stumpf, W.-K. Wong, M. M. Burnett, S. Perona, A. Ko and I. Oberst. ‘Why-Oriented End-User Debugging of Naive Bayes Text Classification’. In: *ACM Trans. Interact. Intell. Syst. (TiiS)* 1.1 (2011), p. 2.

[137] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao et al. ‘Interpretability of deep learning models: a survey of results’. In: *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE. 2017, pp. 1–6.

[138] A. Datta, S. Sen and Y. Zick. ‘Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems’. In: *Proc. 2016 IEEE Symp. Secur. Priv. (SP 2016)*. IEEE, 2016, pp. 598–617. ISBN: 9781509008247. DOI: 10.1109/SP.2016.42.

[139] Tao Lei, Regina Barzilay and Tommi S Jaakkola. 'Rationalizing Neural Predictions'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 2016, pp. 107–117.

[140] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. 'Model-agnostic interpretability of machine learning'. In: *Proceedings of the International Conference on Machine Learning, Workshop on Human Interpretability in Machine Learning*. 2016.

[141] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra. 'Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization'. In: *NIPS 2016 Workshop Interpretable Machine Learning Complex Systems*. 2016.

[142] G. Montavon, S. Lapuschkin, A. Binder, W. Samek and K. R. Müller. 'Explaining Nonlinear Classification Decisions with Deep Taylor Decomposition'. In: *Pattern Recognition* 65 (2017), pp. 211–222.

[143] M. Sundararajan, A. Taly and Q. Yan. 'Axiomatic Attribution for Deep Networks'. In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017.

[144] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen and S. Sclaroff. 'Top-Down Neural Attention by Excitation Backprop'. In: *International Journal of Computer Vision* (2017), pp. 1–19.

[145] A. Nguyen, A. Dosovitskiy, J. Yosinski, T. Brox and J. Clune. 'Synthesizing the Preferred Inputs for Neurons in Neural Networks via Deep Generator Networks'. In: *Adv. Neural Inf. Process. Syst.* 29 (2016).

[146] Upol Ehsan, Brent Harrison, Larry Chan and Mark O Riedl. 'Rationalization: A neural machine translation approach to generating natural language explanations'. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 81–87.

[147] R. Krishnan, G. Sivakumar and P. Bhattacharya. 'Extracting Decision Trees From Trained Neural Networks'. In: *Pattern Recognit.* 32 (1999), pp. 1999–2009. DOI: 10.1145/775047.775113.

[148] J. J. Thiagarajan, B. Kailkhura, P. Sattigeri and K. N. Ramamurthy. 'TreeView: Peeking into Deep Neural Networks Via Feature-Space Partitioning'. In: *NIPS 2016 Workshop Interpretable Machine Learning Complex Systems*. 2016.

[149] Y. Zhou and G. Hooker. 'Interpreting Models via Single Tree Approximation'. In: *arXiv preprint arXiv:1610.09036* (2016). arXiv: 1610.09036.

[150] Daniel Hein, Steffen Udluft and Thomas A Runkler. ‘Interpretable policies for reinforcement learning by genetic programming’. In: *Engineering Applications of Artificial Intelligence* 76 (2018), pp. 158–169.

[151] D. M. Malioutov, K. R. Varshney, A. Emad and S. Dash. ‘Learning Interpretable Classification Rules with Boolean Compressed Sensing’. In: *Transparent Data Mining for Big and Small Data* 32 (2017). DOI: 10.1007/978-3-319-54024-5.

[152] N. Puri, P. Gupta, P. Agarwal, S. Verma and B. Krishnamurthy. ‘MAGIX: Model Agnostic Globally Interpretable Explanations’. In: *Proceedings of the International Conference on Learning Representations, Debugging Machine Learning Models workshop*. 2017.

[153] T. Wang, C. Rudin, F. Velez-Doshi, Y. Liu, E. Klampfl and P. Macneille. ‘Bayesian Rule Sets for Interpretable Classification’. In: *Proc. IEEE Int. Conf. Data Min. (ICDM)*. IEEE, 2017, pp. 1269–1274. ISBN: 9781509054725. DOI: 10.1109/ICDM.2016.130.

[154] Ethan Elenberg, Alexandros G Dimakis, Moran Feldman and Amin Karbasi. ‘Streaming weak submodularity: Interpreting neural networks on the fly’. In: *Advances in Neural Information Processing Systems* 30 (2017).

[155] S. Lundberg and S.-I. Lee. ‘An unexpected unity among methods for interpreting model predictions’. In: *Proceedings of the Conference Neural Information Processing Systems*. 2016.

[156] M. Pacer and T. Lombrozo. ‘Ockham’s Razor Cuts to the Root: Simplicity in Causal Explanation’. In: *J. Exp. Psychol. Gen.* 146.12 (2017), pp. 1761–1780. DOI: 10.1037/xge0000318.

[157] J. Huysmans, K. Dejaeger, C. Mues, J. Vanthienen and B. Baesens. ‘An Empirical Evaluation of the Comprehensibility of Decision Table, Tree and Rule Based Predictive Models’. In: *Decis. Support Syst.* 51.1 (2011), pp. 141–154. DOI: 10.1016/j.dss.2010.12.003.

[158] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam and Payel Das. ‘Explanations based on the missing: Towards contrastive explanations with pertinent negatives’. In: *Advances in neural information processing systems* 31 (2018).

[159] M. W. Craven and J. W. Shavlik. *Rule Extraction: Where Do We Go from Here?* Tech. rep. University of Wisconsin Machine Learning Research Group, 1999.

[160] D. Dua and E. Karra Taniskidou. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml.

[161] Abhinav Verma, Vijayaraghavan Murali, Rishabh Singh, Pushmeet Kohli and Swarat Chaudhuri. 'Programmatically interpretable reinforcement learning'. In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5045–5054.

[162] Tianmin Shu, Caiming Xiong and Richard Socher. 'Hierarchical and Interpretable Skill Acquisition in Multi-task Reinforcement Learning'. In: *International Conference on Learning Representations*. 2018.

[163] Abhijit Gosavi. 'Reinforcement learning: A tutorial survey and recent advances'. In: *INFORMS Journal on Computing* 21.2 (2009), pp. 178–192.

[164] Alexander A Sherstov and Peter Stone. 'Improving action selection in MDP's via knowledge transfer'. In: *AAAI*. Vol. 5. 2005, pp. 1024–1029.

[165] Yuan Wu, Xun Yao, Giacomo Vespasiani, Antonio Nicolucci, Yajie Dong, Joey Kwong, Ling Li, Xin Sun, Haoming Tian and Sheyu Li. 'Mobile app-based interventions to support diabetes self-management: a systematic review of randomized controlled trials to identify functions associated with glycemic efficacy'. In: *JMIR mHealth and uHealth* 5.3 (2017), e35.

[166] Indranil Bose and Radha K Mahapatra. 'Business data mining; a machine learning perspective'. In: *Information & management* 39.3 (2001), pp. 211–225.

[167] Maxwell W Libbrecht and William Stafford Noble. 'Machine learning applications in genetics and genomics'. In: *Nature Reviews Genetics* 16.6 (2015), p. 321.

[168] James Pita, Milind Tambe, Chris Kiekintveld, Shane Cullen and Erin Steigerwald. 'GUARDS: game theoretic security allocation on a national scale'. In: *The 10th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems. 2011, pp. 37–44.

[169] Jurriaan van Diggelen, Hans van den Broek, Jan Maarten Schraagen and Jasper van der Waa. 'An intelligent operator support system for dynamic positioning'. In: *International Conference on Applied Human Factors and Ergonomics*. Springer. 2017, pp. 48–59.

[170] Federico Cabitza, Raffaele Rasoini and Gian Franco Gensini. 'Unintended consequences of machine learning in medicine'. In: *Jama* 318.6 (2017), pp. 517–518.

[171] Jenna Burrell. ‘How the machine ‘thinks’: Understanding opacity in machine learning algorithms’. In: *Big Data & Society* 3.1 (2016).

[172] Been Kim, Elena Glassman, Brittney Johnson and Julie Shah. *iBCM: Interactive Bayesian case model empowering humans via intuitive interaction*. Tech. rep. MIT-CSAIL-TR-2015-010, 2015.

[173] Greg Ridgeway, David Madigan, Thomas Richardson and John O’Kane. ‘Interpretable Boosted Naïve Bayes Classification.’ In: *KDD*. 1998, pp. 101–104.

[174] Andreas Holzinger, André Carrington and Heimo Müller. ‘Measuring the quality of explanations: the system causability scale (SCS)’. In: *KI-Künstliche Intelligenz* 34.2 (2020), pp. 193–198.

[175] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw and Al Underbrink. ‘Trust in automation’. In: *IEEE Intelligent Systems* 28.1 (2013), pp. 84–88.

[176] Elisabeth W Fitzhugh, Robert R Hoffman and Janet E Miller. *Active trust management*. Ashgate, 2011.

[177] Marvin S Cohen, Raja Parasuraman and Jared T Freeman. ‘Trust in decision aids: A model and its training implications’. In: *in Proc. Command and Control Research and Technology Symp*. Citeseer. 1998.

[178] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter and Lalana Kagal. ‘Explaining explanations: An overview of interpretability of machine learning’. In: *IEEE 5th International Conference on data science and advanced analytics*. IEEE. 2018, pp. 80–89.

[179] Bryce Goodman and Seth Flaxman. ‘European Union regulations on algorithmic decision-making and a “right to explanation"’. In: *AI magazine* 38.3 (2016), pp. 50–57.

[180] Jianlong Zhou and Fang Chen. ‘2D Transparency Space; Bring Domain Users and Machine Learning Experts Together’. In: *Human and Machine Learning*. Springer, 2018, pp. 3–19.

[181] Indre Zliobaite. ‘A survey on measuring indirect discrimination in machine learning’. In: *arXiv preprint, arXiv:1511.00148* (2015).

[182] David Landsbergen, David H Coursey, Stephen Loveless and RF Shangraw Jr. ‘Decision quality, confidence, and commitment with expert systems: An experimental study’. In: *Journal of Public Administration Research and Theory* 7.1 (1997), pp. 131–158.

[183] Donald Waterman. *A guide to expert systems*. Addison-Wesley Pub. Co., Reading, MA, 1986.

[184] Peter Walley. 'Measures of uncertainty in expert systems'. In: *Artificial intelligence* 83.1 (1996), pp. 1–58.

[185] G M Foody. 'Local characterization of thematic classification accuracy through spatially constrained confusion matrices'. In: *International Journal of Remote Sensing* 26.6 (Apr. 2005), pp. 1217–1228.

[186] Nicolas Papernot and Patrick McDaniel. 'Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning'. In: *arXiv preprint arXiv:1803.04765* (2018).

[187] John Platt et al. 'Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods'. In: *Advances in large margin classifiers* 10.3 (June 1999), pp. 61–74.

[188] Michael E. Tipping. 'The Relevance Vector Machine'. In: *Advances in Neural Information Processing Systems (NIPS' 2000)*. 2000, pp. 652–658.

[189] Siddhartha Bhattacharyya. 'Confidence in predictions from random tree ensembles'. In: *Knowledge and information systems* 35.2 (2013), pp. 391–410.

[190] Vincent Labatut and Hocine Cherifi. 'Evaluation of performance measures for classifiers comparison'. In: *Ubiquitous Computing and Communication Journal* 6 (2011), pp. 21–34.

[191] Hugo Zaragoza and Florence d'Alché-Buc. 'Confidence measures for neural network classifiers'. In: *Proceedings of the Seventh Int. Conf. Information Processing and Management of Uncertainty in Knowlegde Based Systems*. 1998.

[192] Irene Sturm, Sebastian Lapuschkin, Wojciech Samek and Klaus-Robert Müller. 'Interpretable deep neural networks for single-trial EEG classification'. In: *Journal of neuroscience methods* 274 (2016), pp. 141–145.

[193] Anh Nguyen, Jason Yosinski and Jeff Clune. 'Deep neural networks are easily fooled: High confidence predictions for unrecognizable images'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 427–436.

[194] Ian J Goodfellow, Jonathon Shlens and Christian Szegedy. 'Explaining and harnessing adversarial examples'. In: *Proceedings of the International Conference on Learning Representations*.

[195] Bianca Zadrozny and Charles Elkan. ‘Transforming classifier scores into accurate multiclass probability estimates’. In: *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2002, pp. 694–699.

[196] Bianca Zadrozny and Charles Elkan. ‘Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers’. In: 1 (2001), pp. 609–616.

[197] Cheng-Lin Liu, Hongwei Hao and Hiroshi Sako. ‘Confidence transformation for combining classifiers’. In: *pattern analysis and applications* 7.1 (2004), pp. 2–17.

[198] Hongwei Hao, Cheng-Lin Liu, Hiroshi Sako et al. ‘Confidence Evaluation for Combining Diverse Classifiers.’ In: *ICDAR*. Vol. 3. 2003, pp. 760–765.

[199] Alexandru Niculescu-Mizil and Rich Caruana. ‘Predicting good probabilities with supervised learning’. In: *Proceedings of the 22nd international conference on Machine learning*. ACM. 2005, pp. 625–632.

[200] Irina Rish. ‘An empirical study of the naive Bayes classifier’. In: *IJCAI 2001 workshop on empirical methods in artificial intelligence*. 2001, pp. 41–46.

[201] Yarin Gal and Zoubin Ghahramani. ‘A theoretically grounded application of dropout in recurrent neural networks’. In: *Advances in neural information processing systems*. 2016, pp. 1019–1027.

[202] Meire Fortunato, Charles Blundell and Oriol Vinyals. ‘Bayesian recurrent neural networks’. In: *Proceedings of the Conference on Neural Information Processing Systems, Women in Machine Learning Workshop*. 2017.

[203] Alex Graves. ‘Practical variational inference for neural networks’. In: *Advances in neural information processing systems*. 2011, pp. 2348–2356.

[204] John Paisley, David M Blei and Michael I Jordan. ‘Variational Bayesian Inference with Stochastic Search’. In: *Proceedings of the 29th International Conference on Machine Learning*. 2012.

[205] Alexander Pollatsek, Arnold D. Well, Clifford Konold, Pamela Hardiman and George Cobb. ‘Understanding conditional probabilities’. In: *Organizational Behavior and Human Decision Processes* 40.2 (1987), pp. 255–269.

[206] Jonathan Evans, Simon Handley and David Over. ‘Conditionals and conditional probability.’ In: *Experimental Psychology: Learning, Memory, and Cognition* 29.2 (2003), p. 321.

[207] Zengchang Qin. ‘Naive Bayes classification given probability estimation trees’. In: *2006 5th International Conference on Machine Learning and Applications (ICMLA’06)*. IEEE. 2006, pp. 34–42.

[208] Robi Polikar. ‘Ensemble based systems in decision making’. In: *IEEE Circuits and systems magazine* 6.3 (2006), pp. 21–45.

[209] Merijn Van Erp, Louis Vuurpijl and Lambert Schomaker. ‘An overview and comparison of voting methods for pattern recognition’. In: *Proceedings Eighth International Workshop on Frontiers in Handwriting Recognition*. IEEE. 2002, pp. 195–200.

[210] Norbert Tóth and Béla Pataki. ‘Classification confidence weighted majority voting using decision tree classifiers’. In: *International Journal of Intelligent Computing and Cybernetics* 1.2 (2008), pp. 169–192.

[211] Peter Stone and Manuela Veloso. ‘Using decision tree confidence factors for multiagent control’. In: *Robot Soccer World Cup*. Springer. 1997, pp. 99–111.

[212] Raquel Florez-Lopez and Juan Manuel Ramon-Jeronimo. ‘Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment. A correlated-adjusted decision forest proposal’. In: *Expert Systems with Applications* 42.13 (2015), pp. 5737–5753.

[213] Christopher G. Atkeson, Andrew W. Moore and Stefan Schaal. ‘Locally Weighted Learning’. In: *Artificial Intelligence Review* 11.June (1997), pp. 11–73.

[214] Evelyn Fix and Joseph L Hodges Jr. *Discriminatory analysis-nonparametric discrimination: consistency properties*. Tech. rep. California Univ Berkeley, 1951.

[215] Dietrich Wettschereck, David W Aha and Takao Mohri. ‘A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms’. In: *Artificial Intelligence Review* 11.1-5 (1997), pp. 273–314.

[216] Roger C Schank, Alex Kass and Christopher K Riesbeck. *Inside case-based explanation*. Psychology Press, 2014.

[217] Gerard P Hodgkinson, Janice Langan-Fox and Eugene Sadler-Smith. ‘Intuition: A fundamental bridging construct in the behavioural sciences’. In: *British Journal of Psychology* 99.1 (2008), pp. 1–27.

[218] Christian Harteis and Stephen Billett. ‘Intuitive expertise: Theories and empirical evidence’. In: *Educational Research Review* 9 (2013), pp. 145–157.

[219] Alex A Freitas. ‘Comprehensible classification models: a position paper’. In: *ACM SIGKDD explorations newsletter* 15.1 (2014), pp. 1–10.

[220] Dónal Doyle, Alexey Tsymbal and Pádraig Cunningham. *A review of explanation and explanation in case-based reasoning*. Tech. rep. Trinity College Dublin, Department of Computer Science, 2003.

[221] Susan F McLean. ‘Case-based learning and its application in medical and health-care fields: a review of worldwide literature’. In: *Journal of Medical Education and Curricular Development* 3 (2016), JMECD–S20377.

[222] Almir Olivette Artero, Maria Cristina Ferreira de Oliveira and Haim Levkowitz. ‘Uncovering clusters in crowded parallel coordinates visualizations’. In: *IEEE Symposium on Information Visualization*. IEEE. 2004, pp. 81–88.

[223] Amit Mandelbaum and Daphna Weinshall. ‘Distance-based Confidence Score for Neural Network Classifiers’. MA thesis. Hebrew University of Jerusalem, 2017.

[224] Akshayvarun Subramanya, Suraj Srinivas and R Venkatesh Babu. ‘Confidence estimation in Deep Neural networks via density modelling’. In: *Proceedings of the International Conference on Multimedia and Expo*. 2017.

[225] Klaus Hechenbichler and Klaus Schliep. ‘Weighted k-nearest-neighbor techniques and ordinal classification’. In: *Discussion Paper 399, SFB386* (2004).

[226] Sahibsingh A Dudani. ‘The distance-weighted k-nearest-neighbor rule’. In: *IEEE Transactions on Systems, Man, and Cybernetics* (1976), pp. 325–327.

[227] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay. ‘Scikit-learn: Machine Learning in Python’. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[228] Fevzi Alimoglu, Ethem Alpaydin and Yagmur Denizhan. ‘Combining multiple classifiers for pen-based handwritten digit recognition’. MA thesis. Institute of Graduate Studies in Science and Engineering, Bogazici University, 1996.

[229] Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee and Victor Froelicher. 'International application of a new probability algorithm for the diagnosis of coronary artery disease'. In: *The American journal of cardiology* 64.5 (1989), pp. 304–310.

[230] Jasper van der Waa, Jurriaan van Diggelen, Mark A Neerincx and Stephan Raaijmakers. 'ICM: An Intuitive Model Independent and Accurate Certainty Measure for Machine Learning.' In: *ICAART (2)*. 2018, pp. 314–321.

[231] Glenn Shafer and Vladimir Vovk. 'A tutorial on conformal prediction'. In: *Journal of Machine Learning Research* 9.Mar (2008), pp. 371–421.

[232] Ulf Johansson, Henrik Linusson, Tuve Löfström and Henrik Boström. 'Interpretable regression trees using conformal prediction'. In: *Expert systems with applications* 97 (2018), pp. 394–404.

[233] Thomas Gilovich, Dale Griffin and Daniel Kahneman. *Heuristics and biases: The psychology of intuitive judgment*. Cambridge university press, 2002.

[234] Lisa Legault. 'The need for autonomy'. In: *Encyclopedia of Personality and Individual Differences, Springer, New York, NY* (2016), pp. 1120–1122.

[235] Donghee Shin. 'The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI'. In: *International Journal of Human-Computer Studies* 146 (2021), p. 102551.

[236] Mark A Neerincx, Jasper van der Waa, Frank Kaptein and Jurriaan van Diggelen. 'Using perceptual and cognitive explanations for enhanced human-agent team performance'. In: *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer. 2018, pp. 204–214.

[237] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers and Mark Neerincx. 'Evaluating XAI: A comparison of rule-based and example-based explanations'. In: *Artificial Intelligence* 291 (2021), p. 103404.

[238] Berk Ustun, Alexander Spangher and Yang Liu. 'Actionable recourse in linear classification'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 10–19.

[239] Bilal Hmoud, Varallyai Laszlo et al. 'Will artificial intelligence take over human resources recruitment and selection'. In: *Network Intelligence Studies* 7.13 (2019), pp. 21–30.

[240] Thomas Ploug and Søren Holm. 'The four dimensions of contestable AI diagnostics-A patient-centric approach to explainable AI'. In: *Artificial Intelligence in Medicine* 107 (2020), p. 101901.

[241] Suresh Venkatasubramanian and Mark Alfano. 'The philosophical basis of algorithmic recourse'. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 284–293.

[242] Joel Walmsley. 'Artificial intelligence and the value of transparency'. In: *AI & SOCIETY* 36.2 (2021), pp. 585–595.

[243] Deirdre K Mulligan, Daniel Kluttz and Nitin Kohli. 'Shaping our tools: Contestability as a means to promote responsible algorithmic decision making in the professions'. In: *Available at SSRN 3311894* (2019).

[244] Jean-François Boulicaut and Jérémy Besson. 'Actionability and formal concepts: A data mining perspective'. In: *International Conference on Formal Concept Analysis*. Springer. 2008, pp. 14–31.

[245] European Parlement. *EU guidelines on ethics in artificial intelligence: Context and implementation*. 2019. URL: https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2019)640163. (accessed: 18-8-2022).

[246] Enid Mumford. 'Sociotechnical systems design: Evolving theory and practice'. In: *Computers and democracy* (1987).

[247] Eric L Trist. *The evolution of socio-technical systems*. Vol. 2. Ontario Quality of Working Life Centre Toronto, 1981.

[248] M Georgeff and A Rao. 'Modeling rational agents within a BDI-architecture'. In: *Proc. 2nd Int. Conf. on Knowledge Representation and Reasoning (KR'91). Morgan Kaufmann*. of. 1991, pp. 473–484.

[249] Marlo Vieira dos Santos Souza. 'Choices that make you change your mind: a dynamic epistemic logic approach to the semantics of BDI agent programming languages'. PhD thesis. Universidade Federal do Rio Grande do Sul, 2016.

[250] Catholijn M Jonker, M Birna Van Riemsdijk and Bas Vermeulen. 'Shared mental models'. In: *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. Springer. 2010, pp. 132–151.

[251] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie and Peter Flach. 'FACE: Feasible and actionable counterfactual explanations'. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 2020, pp. 344–350.

[252] Divyat Mahajan, Chenhao Tan and Amit Sharma. 'Preserving causal constraints in counterfactual explanations for machine learning classifiers'. In: *Proceedings of the NeurIPS Workshop on Do the right thing: Machine learning and Causal Inference for improved decision making*. 2019.

[253] Susanne Dandl, Christoph Molnar, Martin Binder and Bernd Bischl. 'Multi-objective counterfactual explanations'. In: *International Conference on Parallel Problem Solving from Nature*. Springer. 2020, pp. 448–469.

[254] T. A. Schoonderwoerd, W. Jorritsma, M. A. Neerincx and K. van den Bosch. 'Human-Centered XAI: Developing Design Patterns for Explanations of Clinical Decision Support Systems'. In: *International Journal of Human-Computer Studies* (2021).

[255] Yanou Ramon, David Martens, Foster Provost and Theodoros Evgeniou. 'A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C'. In: *Advances in Data Analysis and Classification* 14.4 (2020), pp. 801–819.

[256] Shubham Sharma, Jette Henderson and Joydeep Ghosh. 'Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models'. In: *arXiv preprint arXiv:1905.07857* (2019).

[257] Kentaro Kanamori, Takuya Takagi, Ken Kobayashi and Hiroki Arimura. 'DACE: Distribution-Aware Counterfactual Explanation by Mixed-Integer Linear Optimization.' In: *Proceedings of the International Joint Conference on Artificial Intelligence*. 2020, pp. 2855–2862.

[258] Furui Cheng, Yao Ming and Huamin Qu. 'DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models'. In: *IEEE Transactions on Visualization and Computer Graphics* (2020).

[259] Oscar Gomez, Steffen Holter, Jun Yuan and Enrico Bertini. 'ViCE: visual counterfactual explanations for machine learning models'. In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 2020, pp. 531–535.

[260] Azin Ghazimatin, Oana Balalau, Rishiraj Saha Roy and Gerhard Weikum. 'PRINCE: Provider-side interpretability with counterfactual explanations in recommender systems'. In: *Proceedings of the 13th International Conference on Web Search and Data Mining*. 2020, pp. 196–204.

[261] Michael T Lash, Qihang Lin, W Nick Street and Jennifer G Robinson. ‘A budget-constrained inverse classification framework for smooth classifiers’. In: *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2017, pp. 1184–1193.

[262] Yatong Chen, Jialu Wang and Yang Liu. ‘Strategic recourse in linear classification’. In: *arXiv e-prints* (2020), arXiv–2011.

[263] Michael T Lash, Qihang Lin, Nick Street, Jennifer G Robinson and Jeffrey Ohlmann. ‘Generalized inverse classification’. In: *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM. 2017, pp. 162–170.

[264] Shusen Liu, Bhavya Kailkhura, Jize Zhang, Anna M Hiszpanski, Emily Robertson, Donald Loveland and T Han. ‘Explainable Deep Learning for Uncovering Actionable Scientific Insights for Materials Discovery and Design’. In: *ACS Omega* (2022).

[265] Kaivalya Rawal and Himabindu Lakkaraju. ‘Beyond individualized recourse: Interpretable and interactive summaries of actionable recourses’. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 12187–12198.

[266] Michael Downs, Jonathan L Chu, Yaniv Yacoby, Finale Doshi-Velez and Weiwei Pan. ‘CRUDS: Counterfactual Recourse Using Disentangled Subspaces’. In: *ICML Workshop on Human Interpretability in Machine Learning*. 2020.

[267] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim and Joydeep Ghosh. ‘Towards realistic individual recourse and actionable explanations in black-box decision making systems’. In: *arXiv preprint arXiv:1907.09615* (2019).

[268] Ronal Singh, Paul Dourish, Piers Howe, Tim Miller, Liz Sonenberg, Eduardo Velloso and Frank Vetere. ‘Directive explanations for actionable explainability in machine learning applications’. In: *arXiv preprint arXiv:2102.02671* (2021).

[269] Ramaravind K Mothilal, Amit Sharma and Chenhao Tan. ‘Explaining machine learning classifiers through diverse counterfactual explanations’. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 607–617.

[270] Gabriele Tolomei, Fabrizio Silvestri, Andrew Haines and Mounia Lalmas. 'Interpretable predictions of tree-based ensembles via actionable feature tweaking'. In: *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 2017, pp. 465–474.

[271] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas and Jimbo Wilson. 'The what-if tool: Interactive probing of machine learning models'. In: *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 56–65.

[272] Xin Zhang, Armando Solar-Lezama and Rishabh Singh. 'Interpreting neural network judgments via minimal, stable, and symbolic corrections'. In: *Advances in neural information processing systems* 31 (2018).

[273] Carlos Aguilar-Palacios, Sergio Muñoz-Romero and José luis Rojo-Álvarez. 'Cold-start promotional sales forecasting through gradient boosted-based contrastive explanations'. In: *IEEE Access* 8 (2020), pp. 137574–137586.

[274] André Artelt and Barbara Hammer. 'Efficient computation of counterfactual explanations of LVQ models'. In: *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (2020).

[275] Emre Ates, Burak Aksar, Vitus J Leung and Ayse K Coskun. 'Counterfactual Explanations for Multivariate Time Series'. In: *2021 International Conference on Applied Artificial Intelligence (ICAPAI)*. IEEE. 2021, pp. 1–8.

[276] Matt Chapman-Rounds, Marc-Andre Schulz, Erik Pazos and Konstantinos Georgatzis. 'EMAP: Explanation by Minimal Adversarial Perturbation'. In: *arXiv preprint arXiv:1912.00872* (2019).

[277] Zhicheng Cui, Wenlin Chen, Yujie He and Yixin Chen. 'Optimal action extraction for random forests and boosted trees'. In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 179–188.

[278] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam and Payel Das. 'Explanations based on the missing: Towards contrastive explanations with pertinent negatives'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 592–603.

[279] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh and Stefan Lee. 'Counterfactual visual explanations'. In: *International Conference on Machine Learning*. PMLR. 2019, pp. 2376–2384.

[280] Rory Mc Grath, Luca Costabello, Chan Le Van, Paul Sweeney, Farbod Kamiab, Zhao Shen and Freddy Lecue. ‘Interpretable Credit Application Predictions With Counterfactual Explanations’. In: *NIPS Workshop on Challenges and Opportunities for AI in Financial Services: The Impact of Fairness, Explainability, Accuracy, and Privacy*. 2018.

[281] Amir-Hossein Karimi, Bernhard Schölkopf and Isabel Valera. ‘Algorithmic recourse: from counterfactual explanations to interventions’. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 353–362.

[282] Ana Lucic, Harrie Oosterhuis, Hinda Haned and Maarten de Rijke. ‘FOCUS: Flexible optimizable counterfactual explanations for tree ensembles’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 5. 2022, pp. 5313–5322.

[283] David Martens and Foster Provost. ‘Explaining data-driven document classifications’. In: *Mis Quarterly* 38.1 (2014), pp. 73–100.

[284] Jonathan Moore, Nils Hammerla and Chris Watkins. ‘Explaining deep learning models with constrained adversarial examples’. In: *Pacific Rim International Conference on Artificial Intelligence*. Springer. 2019, pp. 43–56.

[285] Martin Pawelczyk, Klaus Broelemann and Gjergji Kasneci. ‘On counterfactual explanations under predictive multiplicity’. In: *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 809–818.

[286] Alexis Ross, Himabindu Lakkaraju and Osbert Bastani. ‘Ensuring Actionable Recourse via Adversarial Training’. In: *arXiv preprint arXiv:2011.06146* (2020).

[287] Masoud Hashemi and Ali Fathi. ‘PermuteAttack: Counterfactual Explanation of Machine Learning Credit Scorecards’. In: *arXiv preprint arXiv:2008.10138* (2020).

[288] Hong-Gyu Jung, Sin-Han Kang, Hee-Dong Kim, Dong-Ok Won and Seong-Whan Lee. ‘Counterfactual explanation based on gradual construction for deep networks’. In: *Pattern Recognition* (2022), p. 108958.

[289] Arnaud Van Looveren and Janis Klaise. ‘Interpretable counterfactual explanations guided by prototypes’. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2021, pp. 650–665.

[290] Ana Lucic, Hinda Haned and Maarten de Rijke. 'Why does my model fail? contrastive local explanations for retail forecasting'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 90–98.

[291] Ezzeldin Tahoun and Andre Kassis. 'Beyond Explanations: Recourse via Actionable Interpretability'. In: *Proceedings of the International Conference on Neural Information Processing Systems*. 2020, pp. 12187–12198.

[292] Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira and David Martens. 'Explainable image classification with evidence counterfactual'. In: *Pattern Analysis and Applications* (2022), pp. 1–21.

[293] Julius von Kügelgen, Amir-Hossein Karimi, Umang Bhatt, Isabel Valera, Adrian Weller and Bernhard Schölkopf. 'On the fairness of causal algorithmic recourse'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 9. 2022, pp. 9584–9594.

[294] Pei Wang and Nuno Vasconcelos. 'SCOUT: Self-aware Discriminant Counterfactual Explanations'. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8981–8990.

[295] Mark T Keane and Barry Smyth. 'Good counterfactuals and where to find them: A case-based technique for generating counterfactuals for explainable ai (xai)'. In: *International Conference on Case-Based Reasoning*. Springer. 2020, pp. 163–178.

[296] Vincent Ballet, Jonathan Aigrain, Thibault Laugel, Pascal Frossard, Marcin Detyniecki et al. 'Imperceptible Adversarial Attacks on Tabular Data'. In: *Proceedings of the NeurIPS Workshop on Robust AI in Financial Services: Data, Fairness, Explainability, Trustworthiness and Privacy*. 2019.

[297] Carlos Fernández-Loría, Foster Provost and Xintian Han. 'Explaining data-driven decisions made by AI systems: the counterfactual approach'. In: *arXiv preprint arXiv:2001.07417* (2020).

[298] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini and Fosca Giannotti. 'Local rule-based explanations of black box decision systems. arXiv 2018'. In: *arXiv preprint arXiv:1805.10820* (2021).

[299] Riccardo Guidotti, Anna Monreale, Stan Matwin and Dino Pedreschi. ‘Black box explanation by learning image exemplars in the latent feature space’. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2019, pp. 189–205.

[300] Amir-Hossein Karimi, Gilles Barthe, Borja Balle and Isabel Valera. ‘Model-agnostic counterfactual explanations for consequential decisions’. In: *International Conference on Artificial Intelligence and Statistics*. 2020, pp. 895–905.

[301] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan and Weng-Keen Wong. ‘Too much, too little, or just right? Ways explanations impact end users’ mental models’. In: *2013 IEEE Symposium on visual languages and human centric computing*. IEEE. 2013, pp. 3–10.

[302] Martin Pawelczyk, Klaus Broelemann and Gjergji Kasneci. ‘Learning model-agnostic counterfactual explanations for tabular data’. In: *Proceedings of The Web Conference 2020*. 2020, pp. 3126–3132.

[303] Adam White and Artur d’Avila Garcez. ‘Measurable Counterfactual Local Explanations for Any Classifier’. In: *Proceedings of the European Conference on Artificial Intelligence*. IOS Press, 2020, pp. 2529–2535.

[304] André Artelt and Barbara Hammer. ‘Convex density constraints for computing plausible counterfactual explanations’. In: *International Conference on Artificial Neural Networks*. Springer. 2020, pp. 353–365.

[305] André Artelt and Barbara Hammer. ‘Convex optimization for actionable & plausible counterfactual explanations’. In: *arXiv preprint arXiv:2105.07630* (2021).

[306] Gagan Bansal. ‘Explanatory Dialogs: Towards Actionable, Interactive Explanations’. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 356–357.

[307] Alejandro Barredo-Arrieta and Javier Del Ser. ‘Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples’. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–7.

[308] Leopoldo Bertossi. ‘An ASP-based approach to counterfactual explanations for classification’. In: *International Joint Conference on Rules and Reasoning*. Springer. 2020, pp. 70–81.

[309] Clément Henin and Daniel Le Métayer. 'A multi-layered approach for tailored black-box explanations'. In: *Proceedings of the ICPR'2020 Workshop Explainable AI*. Springer. 2020.

[310] Maxim Kovalev, Lev Utkin, Frank Coolen and Andrei Konstantinov. 'Counterfactual explanation of machine learning survival models'. In: *Informatica* 32.4 (2021), pp. 817–847.

[311] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard and Marcin Detyniecki. 'Comparison-based inverse classification for interpretability in machine learning'. In: *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*. Springer. 2018, pp. 100–111.

[312] Benjamin J Lengerich, Sandeep Konam, Eric P Xing, Stephanie Rosenthal and Manuela M Veloso. 'Visual Explanations for Convolutional Neural Networks via Input Resampling'. In: *arXiv preprint arXiv:1707.09641* (2017).

[313] Daniel Nemirovsky, Nicolas Thiebaut, Ye Xu and Abhishek Gupta. 'Providing Actionable Feedback in Hiring Marketplaces using Generative Adversarial Networks'. In: *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 2021, pp. 1089–1092.

[314] Chris Russell. 'Efficient search for diverse coherent explanations'. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 20–28.

[315] Yunxia Zhao. 'Fast Real-time Counterfactual Explanations'. In: *Proceedings of the International Conference of Machine Learning*. 2020.

[316] Solon Barocas, Andrew D Selbst and Manish Raghavan. 'The hidden assumptions behind counterfactual explanations and principal reasons'. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 80–89.

[317] Amir-Hossein Karimi, Bernhard Sch¨olkopf and Isabel Valera. 'Algorithmic recourse: from counterfactual explanations to interventions'. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 353–362.

[318] André Artelt and Barbara Hammer. 'Efficient computation of counterfactual explanations and counterfactual metrics of prototype-based classifiers'. In: *Neurocomputing* 470 (2022), pp. 304–317.

[319] Martin Lindvall, Jesper Molin and AB Sectra. 'Verification Staircase: a Design Strategy for Actionable Explanations.' In: *ExSS-ATEC@ IUI*. 2020.

[320] Shusen Liu, Bhavya Kailkhura, Donald Loveland and Yong Han. 'Generative counterfactual introspection for explainable deep learning'. In: *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE. 2019, pp. 1–5.

[321] Shubham Rathi. 'Generating counterfactual and contrastive explanations using SHAP'. In: *Proceedings of the International Joint Conference on AI, Workshop on Humanizing AI*. 2019.

[322] Michael Anderson and Susan Leigh Anderson. 'Machine ethics: Creating an ethical intelligent agent'. In: *AI magazine* 28.4 (2007), pp. 15–15.

[323] High Level Expert Group on Artificial Intelligence. *Ethics guidelines for trustworthy AI*. 2019.

[324] James H Moor. 'The nature, importance, and difficulty of machine ethics'. In: *IEEE intelligent systems* 21.4 (2006), pp. 18–21.

[325] Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng and Max Kramer. 'Moral decision making frameworks for artificial intelligence'. In: *31st AAAI Conference on Artificial Intelligence*. 2017.

[326] Filippo Santoni de Sio and Jeroen Van den Hoven. 'Meaningful human control over autonomous systems: a philosophical account'. In: *Frontiers in Robotics and AI* 5 (2018), p. 15.

[327] Kristina Lerman, Chris Jones, Aram Galstyan and Maja J Matarić. 'Analysis of dynamic task allocation in multi-robot systems'. In: *The International Journal of Robotics Research* 25.3 (2006), pp. 225–241.

[328] Jurriaan van Diggelen and Matthew Johnson. 'Team Design Patterns'. In: *Proceedings of the 7th International Conference on Human-Agent Interaction*. ACM. 2019, pp. 118–126.

[329] Wendell Wallach, Colin Allen and Iva Smit. 'Machine morality: bottom-up and top-down approaches for modelling human moral faculties'. In: *Ai & Society* 22.4 (2008), pp. 565–582.

[330] Batya Friedman and David G Hendry. *Value sensitive design: Shaping technology with moral imagination*. Mit Press, 2019.

[331] IEEE Global Initiative et al. *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. 2018.

[332] Michael Anderson and Susan Leigh Anderson. 'GenEth: A general ethical dilemma analyzer'. In: *Paladyn, Journal of Behavioral Robotics* 9.1 (2018), pp. 337–357.

[333] Ritesh Noothigattu, Snehalkumar S Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar and Ariel D Procaccia. ‘A voting-based system for ethical decision making’. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[334] Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell and Anca Dragan. ‘Inverse reward design’. In: *Advances in neural information processing systems*. 2017, pp. 6765–6774.

[335] Thomas Arnold, Daniel Kasenberg and Matthias Scheutz. ‘Value Alignment or Misalignment–What Will Keep Systems Accountable?’ In: *Workshops at the Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[336] Tae Wan Kim, Thomas Donaldson and John Hooker. ‘Grounding Value Alignment with Ethical Principles’. In: *arXiv preprint arXiv:1907.05447* (2019).

[337] Aimee van Wynsberghe and Scott Robbins. ‘Critiquing the reasons for making artificial moral agents’. In: *Science and engineering ethics* 25.3 (2019), pp. 719–735.

[338] Virginia Dignum. *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing, 2019.

[339] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson et al. ‘Machine behaviour’. In: *Nature* 568.7753 (2019), pp. 477–486.

[340] Mark A Neerincx, Jurriaan van Diggelen and Leo van Breda. ‘Interaction design patterns for adaptive human-agent-robot teamwork in high-risk domains’. In: *International conference on engineering psychology and cognitive ergonomics*. Springer. 2016, pp. 211–220.

[341] Axel Schulte, Diana Donath and Douglas S Lange. ‘Design patterns for human-cognitive agent teaming’. In: *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer. 2016, pp. 231–243.

[342] Jurriaan van Diggelen, Mark Neerincx, Marieke Peeters and Jan Maarten Schraagen. ‘Developing effective and resilient human-agent teamwork using team design patterns’. In: *IEEE intelligent systems* 34.2 (2018), pp. 15–24.

[343] David A Abbink, Tom Carlson, Mark Mulder, Joost CF de Winter, Farzad Aminravan, Tricia L Gibo and Erwin R Boer. 'A topology of shared control systems—finding common ground in diversity'. In: *IEEE Transactions on Human-Machine Systems* 48.5 (2018), pp. 509–525.

[344] Fanny Ficuciello, Guglielmo Tamburrini, Alberto Arezzo, Luigi Villani and Bruno Siciliano. 'Autonomy in surgical robots and its meaningful human control'. In: *Paladyn, Journal of Behavioral Robotics* 10.1 (2019), pp. 30–43.

[345] Shane O'Sullivan, Nathalie Nevejans, Colin Allen, Andrew Blyth, Simon Leonard, Ugo Pagallo, Katharina Holzinger, Andreas Holzinger, Mohammed Imran Sajid and Hutan Ashrafian. 'Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery'. In: *The International Journal of Medical Robotics and Computer Assisted Surgery* 15.1 (2019), e1968.

[346] Thomas Tung and Claude H Organ. 'Ethics in surgery: historical perspective'. In: *Archives of Surgery* 135.1 (2000), pp. 10–13.

[347] Guus Beckers, Joris Sijs, Jurriaan van Diggelen, Roelof JE van Dijk, Henri Bouma, Mathijs Lomme, Rutger Hommes, Fieke Hillerstrom, Jasper van der Waa, Anna van Velsen et al. 'Intelligent autonomous vehicles with an extendable knowledge base and meaningful human control'. In: *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies III*. Vol. 11166. International Society for Optics and Photonics. 2019, p. 111660C.

[348] Roger Clarke. 'The regulation of civilian drones' impacts on behavioural privacy'. In: *Computer Law & Security Review* 30.3 (2014), pp. 286–305.

[349] Richard M Thompson. 'Drones in domestic surveillance operations: Fourth amendment implications and legislative responses'. In: Congressional Research Service, Library of Congress. 2012.

[350] Caspar G Chorus. 'Models of moral decision making: Literature review and research agenda for discrete choice analysis'. In: *Journal of choice modelling* 16 (2015), pp. 69–85.

[351] Robert J Sternberg. 'A model for ethical reasoning'. In: *Review of General psychology* 16.4 (2012), pp. 319–326.

[352] Stuart Russell, Sabine Hauert, Russ Altman and Manuela Veloso. 'Ethics of artificial intelligence'. In: *Nature* 521.7553 (2015), pp. 415–416.

[353] Weiyu Wang and Keng Siau. 'Ethical and moral issues with AI: a case study on healthcare robots'. In: *Twenty-fourth Americas conference on information systems. Retrieved from July*. Vol. 16. 2018, p. 2019.

[354] Simeon C Calvert, Daniël D Heikoop, Giulio Mecacci and Bart van Arem. 'A human centric framework for the analysis of automated driving systems based on meaningful human control'. In: *Theoretical Issues in Ergonomics Science* 21.4 (2020), pp. 478–506.

[355] Michael C Horowitz and Paul Scharre. *Meaningful human control in weapon systems: A primer*. Tech. rep. Washington, DC; Center for a New American Security, 2015.

[356] Marieke MM Peeters, Jurriaan van Diggelen, Karel Van Den Bosch, Adelbert Bronkhorst, Mark A Neerincx, Jan Maarten Schraagen and Stephan Raaijmakers. 'Hybrid collective intelligence in a human–AI society'. In: *AI & SOCIETY* (2020), pp. 1–22.

[357] Wendell Wallach and Shannon Vallor. 'Moral Machines: From Value Alignment to Embodied Virtue'. In: *Ethics of Artificial Intelligence*. Oxford University Press, 2020, pp. 383–412.

[358] Rebecca Crootof. 'A Meaningful Floor for Meaningful Human Control'. In: *Temp. Int'l & Comp. LJ* 30 (2016), p. 53.

[359] Article 36. *Key areas for debate on autonomous weapons systems*. https://www.article36.org/wp-content/uploads/2014/05/A36-CCW-May-2014.pdf (accessed: 2020-11-09). 2014.

[360] Ronald Arkin. *Governing lethal behavior in autonomous robots*. CRC Press, 2009.

[361] Judd E Hollander and Brendan G Carr. 'Virtually perfect? Telemedicine for COVID-19'. In: *New England Journal of Medicine* 382.18 (2020), pp. 1679–1681.

[362] Sarmad Sadeghi, Afsaneh Barzi, Navid Sadeghi and Brent King. 'A Bayesian model for triage decision support'. In: *International journal of medical informatics* 75.5 (2006), pp. 403–411.

[363] Michael Boardman and Fiona Butcher. *An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It*. Tech. rep. NATO, 2019.

[364] Merel AC Ekelhof. 'Lifting the Fog of Targeting'. In: *Naval War College Review* 71.3 (2018), pp. 61–95.

[365] Marcello Guarini and Paul Bello. ‘Robotic warfare: some challenges in moving from noncivilian to civilian theaters’. In: *Robot ethics: The ethical and social implications of robotics* 129 (2012), p. 136.

[366] Glen Klein, David D Woods, Jeffrey M Bradshaw, Robert R Hoffman and Paul J Feltovich. ‘Ten challenges for making automation a" team player" in joint human-agent activity’. In: *IEEE Intelligent Systems* 19.6 (2004), pp. 91–95.

[367] Geert-Jan M Kruijff and Miroslav Janıcek. ‘Using Doctrines for Human-Robot Collaboration to Guide Ethical Behavior.’ In: *AAAI fall symposium: robot-human teamwork in dynamic adverse environment*. Citeseer. 2011, pp. 26–33.

[368] Jurriaan van Diggelen, JS Barnhoorn, Marieke MM Peeters, Wessel van Staal, M van Stolk, Bob van der Vecht, Jasper van der Waa and Jan Maarten Schraagen. ‘Pluggable social artificial intelligence for enabling human-agent teaming’. In: *Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems*. 2019.

[369] Michael J Barnes, Jessie Y Chen and Susan Hill. *Humans and autonomy: Implications of shared decision making for military operations*. Tech. rep. US Army Research Laboratory Aberdeen Proving Ground United States, 2017.

[370] D Doran, SC Schulz and TR Besold. ‘What Does Explainable AI Really Mean? A New Conceptualization of Perspectives’. In: *CEUR Workshop Proceedings*. Vol. 2071. CEUR. 2018.

[371] Ze Gong and Yu Zhang. ‘Behavior explanation as intention signaling in human-robot teaming’. In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE. 2018, pp. 1005–1011.

[372] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright and Michael Barnes. ‘Situation awareness-based agent transparency and human-autonomy teaming effectiveness’. In: *Theoretical issues in ergonomics science* 19.3 (2018), pp. 259–282.

[373] Elham Khodabandehloo, Daniele Riboni and Abbas Alimohammadi. ‘HealthXAI: Collaborative and explainable AI for supporting early diagnosis of cognitive decline’. In: *Future Generation Computer Systems* (2020).

[374] Taemie Kim and Pamela Hinds. 'Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction'. In: *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2006, pp. 80–85.

[375] Charles E Billings. *Human-centered aviation automation: Principles and guidelines*. Tech. rep. Washington, DC: NASA, 1996.

[376] Jasper van der Waa, Tjeerd Schoonderwoerd, Jurriaan van Diggelen and Mark Neerincx. 'Interpretable confidence measures for decision support systems'. In: *International Journal of Human-Computer Studies* 144 (2020), p. 102493.

[377] Josua Krause, Adam Perer and Enrico Bertini. 'A user study on the effect of aggregating explanations for interpreting machine learning models'. In: *ACM KDD Workshop on Interactive Data Exploration and Analytics*. 2018.

[378] Juntang Zhuang, Nicha C Dvornek, Xiaoxiao Li, Junlin Yang and James Duncan. 'Decision explanation and feature importance for invertible networks'. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE. 2019, pp. 4235–4239.

[379] Karen Simonyan, Andrea Vedaldi and Andrew Zisserman. 'Deep inside convolutional networks: Visualising image classification models and saliency maps'. In: *arXiv preprint arXiv:1312.6034* (2013).

[380] Scott M Lundberg, Gabriel G Erion and Su-In Lee. 'Consistent individualized feature attribution for tree ensembles'. In: *Proceedings of the ICML Workshop on Human Interpretability in Machine Learning*. 2018.

[381] Erik Štrumbelj and Igor Kononenko. 'Explaining prediction models and individual predictions with feature contributions'. In: *Knowledge and information systems* 41.3 (2014), pp. 647–665.

[382] Marco Tulio Ribeiro, Sameer Singh and Carlos Guestrin. 'Anchors: High-Precision Model-Agnostic Explanations.' In: *AAAI*. Vol. 18. 2018, pp. 1527–1535.

[383] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan and Been Kim. 'The (un) reliability of saliency methods'. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019, pp. 267–280.

[384] Gabriëlle Ras, Marcel van Gerven and Pim Haselager. 'Explanation methods in deep learning: Users, values, concerns and challenges'. In: *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, 2018, pp. 19–36.

[385] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt and Been Kim. 'Sanity checks for saliency maps'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 9505–9515.

[386] David Alvarez Melis and Tommi Jaakkola. 'Towards robust interpretability with self-explaining neural networks'. In: *Advances in Neural Information Processing Systems*. 2018, pp. 7775–7784.

[387] Giles Hooker and Lucas Mentch. 'Please stop permuting features: An explanation and alternatives'. In: *arXiv preprint arXiv:1905.03151* (2019).

[388] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn. 'Bias in random forest variable importance measures: Illustrations, sources and a solution'. In: *BMC bioinformatics* 8.1 (2007), p. 25.

[389] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis. 'Conditional variable importance for random forests'. In: *BMC bioinformatics* 9.1 (2008), p. 307.

[390] Laura Toloşi and Thomas Lengauer. 'Classification with correlated features: unreliability of feature ranking and solutions'. In: *Bioinformatics* 27.14 (2011), pp. 1986–1994.

[391] Botty Dimanov, Umang Bhatt, Mateja Jamnik and Adrian Weller. 'You Shouldn't Trust Me: Learning Models Which Conceal Unfairness From Multiple Explanation Methods.' In: *SafeAI@ AAAI*. 2020, pp. 63–73.

[392] Amirata Ghorbani, Abubakar Abid and James Zou. 'Interpretation of neural networks is fragile'. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 2019, pp. 3681–3688.

[393] Sarthak Jain and Byron C Wallace. 'Attention is not explanation'. In: *Proceedings of the North American Association for Computational Linguistics*. 2019.

[394] Christian Kruschitz and Martin Hitz. 'Human-computer interaction design patterns: structure, methods, and tools'. In: *International Journal of Advanced Software* 3.1 (2010).

[395] Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university press, 1977.

[396] Jasper van der Waa, Jurriaan van Diggelen, Luciano Cavalcante Siebert, Mark Neerincx and Catholijn Jonker. ‘Allocation of moral decision-making in human-agent teams: A pattern approach’. In: *International Conference on Human-Computer Interaction*. Springer. 2020, pp. 203–220.

[397] Colin Allen, Iva Smit and Wendell Wallach. ‘Artificial morality: Top-down, bottom-up, and hybrid approaches’. In: *Ethics and information technology* 7.3 (2005), pp. 149–155.

[398] Jasper van der Waa and Tjalling Haije. *MATRX Software*. https://matrx-software.com/ (accessed 2020-11-09). 2019.

[399] Marsha E Fonteyn, Benjamin Kuipers and Susan J Grobe. ‘A description of think aloud method and protocol analysis’. In: *Qualitative health research* 3.4 (1993), pp. 430–441.

[400] Holzinger Andreas, Carrington André and Müller Heimo. ‘Measuring the Quality of Explanations: The System Causability Scale (SCS): Comparing Human and Machine Explanations’. In: *Kunstliche intelligenz* 34.2 (2020), pp. 193–198.

[401] Mauro Dragoni, Ivan Donadello and Claudio Eccher. ‘Explainable AI meets persuasiveness: Translating reasoning results into behavioral change advice’. In: *Artificial Intelligence in Medicine* 105 (2020), p. 101840.

[402] Nicolás E Díaz Ferreyra, Esma Aïmeur, Hicham Hage, Maritta Heisel and Catherine García van Hoogstraten. ‘Persuasion meets AI: Ethical considerations for the design of social engineering countermeasures’. In: *Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*. 2020.

[403] Oliviero Stock, Marco Guerini and Fabio Pianesi. ‘Ethical dilemmas for adaptive persuasion systems’. In: *Thirtieth AAAI Conference on Artificial Intelligence*. 2016.

[404] Marc Schwartz. ‘Khan academy: The illusion of understanding’. In: *Online Learning Journal* 17.4 (2013).

[405] Pigi Kouki, James Schaffer, Jay Pujara, John O’Donovan and Lise Getoor. ‘Personalized explanations for hybrid recommender systems’. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 2019, pp. 379–390.

[406] Millecamp Martijn, Cristina Conati and Katrien Verbert. ‘“Knowing me, knowing you”: personalized explanations for a music recommender system’. In: *User Modeling and User-Adapted Interaction* (2022), pp. 1–38.

[407] Sarath Sreedharan, Siddharth Srivastava and Subbarao Kambhampati. ‘Using state abstractions to compute personalized contrastive explanations for AI agent behavior’. In: *Artificial Intelligence* 301 (2021), p. 103570.

[408] Tathagata Chakraborti, Sarath Sreedharan and Subbarao Kambhampati. ‘The Emerging Landscape of Explainable Automated Planning & Decision Making.’ In: *IJCAI*. 2020, pp. 4803–4811.

[409] Alnour Alharin, Thanh-Nam Doan and Mina Sartipi. ‘Reinforcement learning interpretation methods: A survey’. In: *IEEE Access* 8 (2020), pp. 171058–171077.

[410] Tatsuya Sakai and Takayuki Nagai. ‘Explainable autonomous robots: a survey and perspective’. In: *Advanced Robotics* (2022), pp. 1–20.

[411] Erika Puiutta and Eric Veith. ‘Explainable reinforcement learning: A survey’. In: *International cross-domain conference for machine learning and knowledge extraction*. Springer. 2020, pp. 77–95.

[412] Zhengxian Lin, Kin-Ho Lam and Alan Fern. ‘Contrastive Explanations for Reinforcement Learning via Embedded Self Predictions’. In: *International Conference on Learning Representations*. 2020.

[413] Sarath Sreedharan, Utkarsh Soni, Mudit Verma, Siddharth Srivastava and Subbarao Kambhampati. ‘Bridging the Gap: Providing Post-Hoc Symbolic Explanations for Sequential Decision-Making Problems with Inscrutable Representations’. In: *International Conference on Learning Representations*. 2021.

[414] Prashan Madumal, Tim Miller, Liz Sonenberg and Frank Vetere. ‘Distal explanations for explainable reinforcement learning agents’. In: *arXiv preprint arXiv:2001.10284* (2020).

[415] David Gunning. ‘Explainable artificial intelligence (xai)’. In: *Defense Advanced Research Projects Agency (DARPA), nd Web* (2017).

[416] Luciano Floridi, Josh Cowls, Thomas C King and Mariarosaria Taddeo. ‘How to design AI for social good: Seven essential factors’. In: *Ethics, Governance, and Policies in Artificial Intelligence*. Springer, 2021, pp. 125–151.

[417] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson and Lance Kaplan. ‘Rapid trust calibration through interpretable and uncertainty-aware AI’. In: *Patterns* 1.4 (2020), p. 100049.

[418] Krishna Prakash Kalyanathaya et al. 'A Literature Review and Research Agenda on Explainable Artificial Intelligence (XAI)'. In: *International Journal of Applied Engineering and Management Letters (IJAEML)* 6.1 (2022), pp. 43–59.

[419] Roman Lukyanenko, Arturo Castellanos, Binny M Samuel, Monica Tremblay and Wolfgang Maass. 'Research Agenda for Basic Explainable AI'. In: *Americas Conference on Information Systems*. 2021, pp. 1–8.

[420] Claudio Sarra. 'Put Dialectics into the Machine: Protection against Automatic-decision-making through a Deeper Understanding of Contestability by Design'. In: *Global Jurist* 20.3 (2020).

[421] Thomas Ploug and Søren Holm. 'Right to Contest AI Diagnostics: Defining Transparency and Explainability Requirements from a Patient's Perspective'. In: *Artificial Intelligence in Medicine*. Springer, 2022, pp. 227–238.

[422] Kars Alfrink, Ianus Keller, Gerd Kortuem and Neelke Doorn. 'Contestable AI by Design: Towards a Framework'. In: *Minds and Machines* (2022), pp. 1–27.

[423] Mark A Neerincx and Jasper Lindenberg. *Situated cognitive engineering for complex task environments*. Ashgate Publishing Limited Aldershot, 2008.

[424] Tim F Paymans, Jasper Lindenberg and Mark Neerincx. 'Usability trade-offs for adaptive user interfaces: ease of use and learnability'. In: *Proceedings of the 9th international conference on Intelligent user interfaces*. 2004, pp. 301–303.

[425] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard and Marcin Detyniecki. 'The Dangers of Post-hoc Interpretability: Unjustified Counterfactual Explanations'. In: *International Joint Conference on Artificial Intelligence*. 2019, pp. 2801–2807.

[426] Sebastian Bordt, Michèle Finck, Eric Raidl and Ulrike von Luxburg. 'Post-Hoc Explanations Fail to Achieve their Purpose in Adversarial Contexts'. In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. 2022.

[427] Henrietta Lyons, Eduardo Velloso and Tim Miller. 'Conceptualising contestability: Perspectives on contesting algorithmic decisions'. In: *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW1 (2021), pp. 1–25.

[428] Wei Xu, Marvin J Dainoff, Liezhong Ge and Zaifeng Gao. 'Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI'. In: *International Journal of Human–Computer Interaction* (2022), pp. 1–25.

[429] Maxwell Szymanski, Martijn Millecamp and Katrien Verbert. 'Visual, textual or hybrid: the effect of user expertise on different explanations'. In: *26th International Conference on Intelligent User Interfaces*. 2021, pp. 109–119.

[430] Aditi Mishra, Utkarsh Soni, Jinbin Huang and Chris Bryan. 'Why? Why not? When? Visual Explanations of Agent Behaviour in Reinforcement Learning'. In: *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*. IEEE. 2022, pp. 111–120.

[431] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel and David C Atkins. 'Designing contestability: Interaction design, machine learning, and mental health'. In: *Proceedings of the 2017 Conference on Designing Interactive Systems*. 2017, pp. 95–99.

[432] Tauseef Ibne Mamun, Robert R Hoffman and Shane T Mueller. 'Collaborative Explainable AI: A non-algorithmic approach to generating explanations of AI'. In: *International Conference on Human-Computer Interaction*. Springer. 2021, pp. 144–150.

[433] European Committee. *A European approach to artificial intelligence*. 2021. URL: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence. (accessed: 18-8-2022).

[434] TNO. *The FATE Project*. 2022. URL: https://appl-ai-tno.nl/projects/fate/. (accessed: 18-8-2022).

[435] TNO. *AI Skills Matching*. 2022. URL: https://appl-ai-tno.nl/projects/skills-matching/. (accessed: 18-8-2022).

[436] TNO. *The AI4Justice Project*. 2021. URL: https://appl-ai-tno.nl/projects/ai4justice/. (accessed: 18-8-2022).

[437] TNO. *DASH: Delegation of Autonomous Systems in Human-Agent teams*. 2021. URL: https://appl-ai-tno.nl/projects/dash/. (accessed: 18-8-2022).

[438] TNO. *The AI Oversight Lab*. 2022. URL: https://appl-ai-tno.nl/projects/ai-oversight-lab. (accessed: 18-8-2022).

[439] TNO. *EU Horizon MOSES (No 861678)*. 2022. URL: https://moses-h2020.eu/. (accessed: 18-8-2022).

[440] Hybrid Intelligence Zwaartekrachtprogramma. *Hybrid Intelligence Centre*. 2022. URL: https://www.hybrid-intelligence-centre.nl/. (accessed: 18-8-2022).

[441] CEE AI Dresden. *Center for Explainable and Efficient AI Technologies*. 2022. URL: https://cee-ai.org/. (accessed: 18-8-2022).

[442] University of Pisa. *The XAI Project*. 2022. URL: https://xai-project.eu/index.html. (accessed: 18-8-2022).

[443] NLAIC. *NLAIC: Human-Centric AI*. 2022. URL: https://nlaic.com/en/building-blocks/human-centric-ai/. (accessed: 18-8-2022).

[444] ELLIS Society. *ELLIS: European Laboratory for Learning and Intelligent Systems*. 2018. URL: https://ellis.eu/. (accessed: 18-8-2022).

[445] CLAIRE AI. *Confederation of Laboratories for AI Research in Europe*. 2018. URL: https://claire-ai.org/. (accessed: 18-8-2022).

ABOUT THE AUTHOR

Jasper van der Waa, born 1991 in the Netherlands, studied Artificial Intelligence at the Radboud University in Nijmegen. After his master's focused on interactive machine learning, he started at TNO in 2016. TNO is the Dutch research institute for applied research, where Jasper's research revolves around optimizing the human-AI interaction by expanding the AI's functionality through the identification of requirements and development of new algorithms and technologies. In 2018 he started with his PhD research at the Technical University of Delft on the topic of explainable AI for improving the human-AI collaboration. His unique position allows him to collaborate closely with companies and government institutes to identify the most pressing issues that can be solved with explainable AI while collaborating with universities to address these issues through research.

He worked with financial institutes (ABN Amro, Achmea), AI companies (Xomnia, Enjin), healthcare institutes (Curium), the Dutch government (Ministry of Defence, Ministry of Internal Affairs, Employee Insurance Agency) and other research institutes (NLR, Marin). He also acted as liaison between the Technical University of Delft and the Ministry of Defence (KMA) on moral value elicitation. Furthermore, he participates in NATO meetings (SCI-335 SRM) and roundtable discussions on the topic of AI in relation to current and future (European) regulations. As of this writing he leads projects that bring human-AI collaboration research to practice. These projects take place in Dutch and European AI-ecosystems, such as NLAIC, the Hybrid Intelligence Centre, and Horizon Europe. Through these projects, he connects applied and fundamental research institutes.

Aside from his collaboration with various parties, he organizes community events such as summer schools, hackathons, and panel discussions that revolve around human-AI interaction. Furthermore, he is the founder of two open-source projects and their communities. The first is MATRX, a tool to accelerate human-machine teaming research by offering a platform to evaluate and test that research in various tasks. MATRX is used in human-AI collaboration experiments and courses about this topic. His newest open-source project is the XAI Toolbox that supports the responsible evaluation of explanation generating methods.

His goal in these efforts is to enable effective and responsible human-AI collaboration and build and sustain communities that share this goal for the good of society.

Jasper van der Waa

10-04-1991 Born in Heerlen, Netherlands.

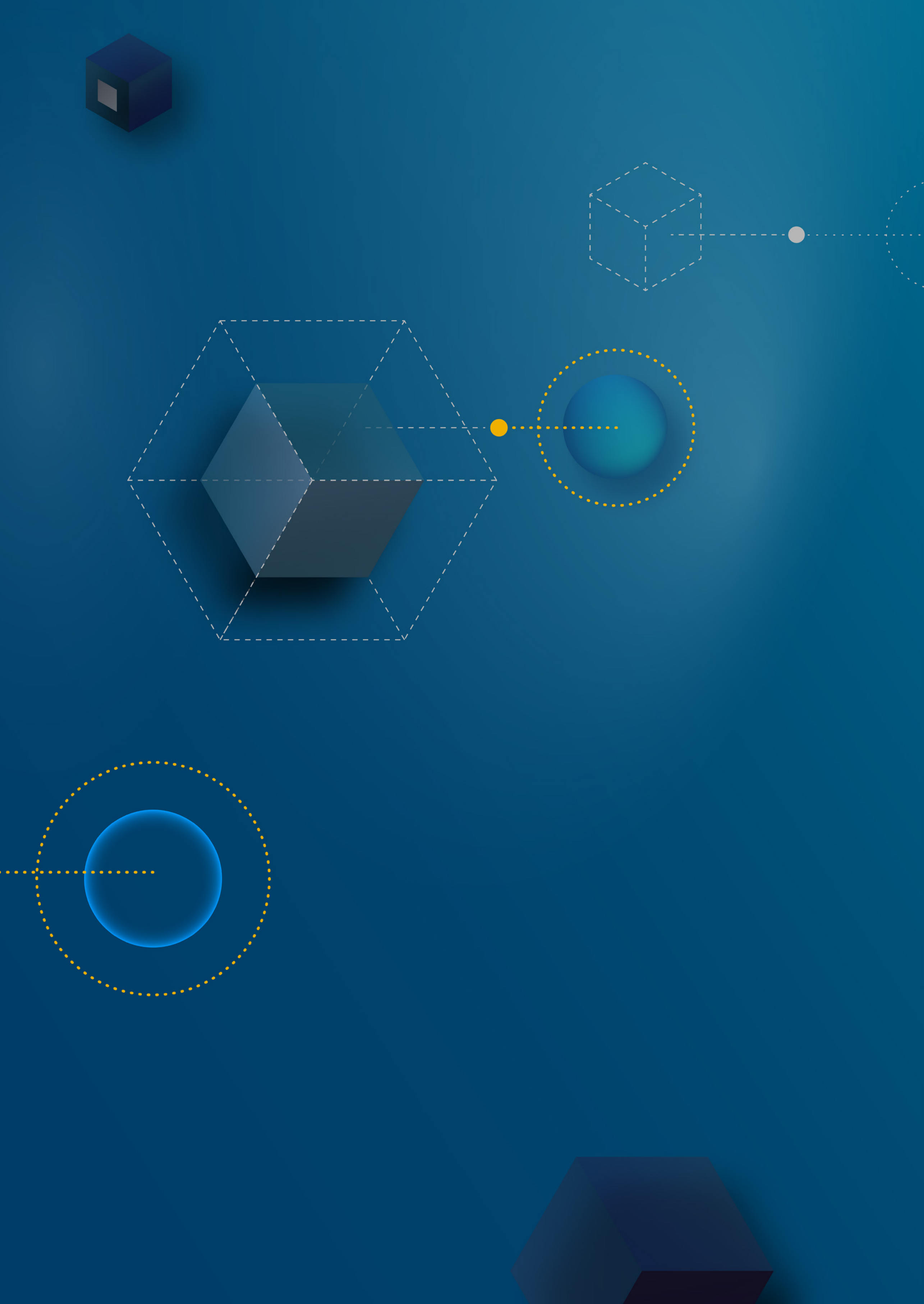
Education

2003–2009	Grammar School Trevianum Scholengroep, Sittard
2009–2013	Undergraduate in Artificial Intelligence Radboud University Nijmegen
2013–2015	Graduate in Artificial Intelligence Radboud University Nijmegen
2016–	Researcher Human–AI interaction TNO, Netherlands
2018–2022	PhD. Artificial Intelligence Technical University of Delft

Thesis:	Explainable AI for Human-AI Collaboration
Promotors:	Prof. dr. M.A. Neerincx Prof. dr. C.M. Jonkers
Supervisor:	Dr. J. van Diggelen

Activities & Merits

2022–	Lead scientist on TNO’s research towards applied and responsible explainable AI for the Dutch military.
2022–	Lead scientist on TNO’s research towards the responsible application of explanations in smart energy systems.
2022–	Project leader and lead scientist on the project Hybrid Intelligence in Practice.
2022–	One of the lead scientists on TNO’s research towards artificial moral agents.
2022	Conference chair of the HHAI conference and hackathon organizer.
2022	Organizer of round table discussions with Dutch companies on their explainable AI needs.
2021–2022	Author on the human-machine teaming theme within the AI/DS Research Agenda for Dutch Ministry of Defense.
2021–2022	Lead scientist on TNO’s explainable AI research for smart energy systems.
2021	Interviewed by the Dutch Internal Affairs for their policy on explainable AI.
2021	Author Human-Machine Teaming Roadmap for the Dutch Defense.
2021	Invited speaker for the Responsible Applications of AI webinar organized by Xomnia.
2021	Participant in NATO Specialist Meeting on Autonomy From A System Perspective (SCI-335 SRM).
2021	Invited speaker for companies (ABN AMRO, Achmea, Enjin) on best-practices for explainable AI.
2020–2022	Task lead in H2020 MOSES on remote operator support within the maritime domain.
2020–2022	Scientific lead on TNO’s research in delegation within human-machine teams.
2020–2022	Lead on TNO’s explainable AI research in the project Appl.AI FATE.
2020	Coauthor of TNO’s PhD student policy.
2019–2021	Founder of MATRX, a simulation and experimentation tool for human-machine teaming research.
2019–2021	Liason of the TNO and ABN AMRO collaboration on applied explainable AI.
2019	Liason between Technical University of Delft, TNO and the Dutch military on moral value elicitation.
2018	Organizer for the first Dutch hackathon on Explainable AI.
2018	Invited speaker on Explainable AI at the BNAIC conference.
2017	Co-organizer of the Human-Robot Interaction Summerschool.



LIST OF PUBLICATIONS

22. Poort, J., Mannucci, T., **van der Waa, J.**, & Shoebi Omrani, P. (under review). An optimum well control using reinforcement learning and policy transfer; application to production optimization and slugging minimization. *Proceedings of the SPE Annual Technical Conference and Exhibition*.

21. **van der Waa, J.**, van Diggelen, J., Neerincx, M. & Jonker. C. (under review). Actionable Explanations for Contestable AI. *Journal of Artificial Intelligence Research*.

20. Adhikari, A., Wenink, E., **van der Waa, J.S.**, Bouter, C., Tolios, I. & Raaijmakers, S. (2022, June). Towards FAIR explainable AI: A standardized ontology for mapping XAI solutions to use cases, explanations, and AI systems. *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*.

19. de Boer, M.H.T., Vethman, S., Bakker, R.M., Adhikari, A., Marcus, M., de Greeff, J., **van der Waa, J.S.**, Schoonderwoerd, T.A.J., Tolios, I., van Zoelen, E.M., Hillerström, F. & Kamphorst. B. (2022, April). The FATE system iterated: Fair, transparent and explainable decision making in a juridical Case. *Proceedings of the AAAI Spring Symposium on Machine Learning and Knowledge Engineering for Hybrid Intelligence*.

18. Kunneman, Y., Alves da Motta-Filho, M. & **van der Waa, J.** (2022, March). Data science for service design: An introductory overview of methods and opportunities. *The Design Journal, 25:2*.

17. **van der Waa, J.**, Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence, 291:103404*.

16. van Diggelen, J., Barnhoorn, J., Post, R., Sijs, J., van der Stap, N., & **van der Waa, J.** (2021, February). Delegation in Human-machine Teaming: Progress, Challenges and Prospects. *Proceedings of the International Conference on Intelligent Human Systems Integration (pp. 10-16). Springer*.

15. **van der Waa, J.**, Verdult, S., van den Bosch, K., van Diggelen, J., Haije, T., van der Stigchel, B., & Cocu, I. (2021). Moral decision making in human-agent teams: Human control and the role of explanations. *Frontiers in Robotics and AI, 8:640647*.

14. **van der Waa, J.**, Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies, 144:102493*.

13. **van der Waa, J.**, van Diggelen, J., Siebert, L. C., Neerincx, M., & Jonker, C. (2020, July). Allocation of moral decision-making in human-agent teams: A pattern approach. *Proceedings of the International Conference on Human-Computer Interaction (pp. 203-220). Springer*.

12. Beckers, G., Sijs, J., van Diggelen, J., van Dijk, R. J., Bouma, H., Lomme, M., Hommes, R., Hillerström, F., **van der Waa, J.**, van Velsen, A., Mannucci, T., Voogd, J., van Staal, W., Veltman, K., Wessels, P. & Huizing, A. (2019, October). Intelligent autonomous vehicles with an extendable knowledge base under meaningful human control. *Counterterrorism, Crime Fighting, Forensics, and Surveillance Technologies. International Society for Optics and Photonics*.

11. Brouwer, A. M., **van der Waa, J.**, & Stokking, H. (2018). BCI to potentially enhance streaming images to a VR headset by predicting head rotation. *Frontiers in human neuroscience, 12:420*.

10. **van der Waa, J.**, van Diggelen, J., van den Bosch, K., & Neerincx, M. (2018, July). Contrastive explanations for reinforcement learning in terms of expected consequences. *Proceedings of the Workshop on Explainable AI, International Joint Conference of Artificial Intelligence*.

9. Neerincx, M., **van der Waa, J.**, Kaptein, F., & van Diggelen, J. (2018, July). Using perceptual and cognitive explanations for enhanced human-agent team performance. *Proceedings of the International Conference on Engineering Psychology and Cognitive Ergonomics (pp. 204-214). Springer*.

8. **van der Waa, J.**, Robeer, M., van Diggelen, J., Brinkhuis, M., & Neerincx, M. (2018, July). Contrastive explanations with local foil trees. *Proceedings of the Workshop on Interpretable Machine Learning, International Joint Conference of Artificial Intelligence*.

7. van der Vecht, B., van Diggelen, J., Peeters, M., Barnhoorn, J., & **van der Waa, J.** (2018, June). SAIL: a social artificial intelligence layer for human-machine teaming. *Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 262-274). Springer*.

6. **van der Waa, J.**, van Diggelen, J., & Neerincx, M. (2018, March). The design and validation of an intuitive confidence measure. *Proceedings of the ACM Workshop On Explainable Smart Systems, International Conference on Intelligent User Interfaces, ACM*.

5. van Diggelen, J., Barnhoorn, J. S., Peeters, M. M., van Staal, W., Stolk, M. L., van der Vecht, B., **van der Waa, J.** & Schraagen, J. M. (2018, February). Pluggable social artificial intelligence for enabling human-agent teaming. *Proceedings of the International Conference on Practical Applications of Agents and Multi-Agent Systems (pp. 361–365). Springer*.

4. **van der Waa, J.**, van Diggelen, J., Neerincx, M. A., & Raaijmakers, S. (2018, January). ICM: An Intuitive Model Independent and Accurate Certainty Measure for Machine Learning. *Proceedings of the International Conference on Agents and Artificial Intelligence, (pp. 314-321), SciTePress*.

3. van Diggelen, J., Looije, R., **van der Waa, J.**, & Neerincx, M. (2017, July). Human robot team development: An operational and technical perspective. *Proceedings of the International Conference on Applied Human Factors and Ergonomics (pp. 293-302). Springer*.

2. van Diggelen, J., van den Broek, H., Schraagen, J. M., & **van der Waa, J.** (2017, July). An intelligent operator support system for dynamic positioning. *Proceedings of the International Conference on Applied Human Factors and Ergonomics (pp. 48-59). Springer*.

1. Brouwer, A. M., **van der Waa, J.** S., Hogervorst, M. A., Cacace, A., & Stokking, H. (2017, March). A feasible BCI in real life: using predicted head rotation to improve HMD imaging. *Proceedings of the ACM Workshop on An Application-oriented Approach to BCI out of the laboratory, ACM*.