University of Zurich[UZH]

Department of Informatics

The Right Thing To Do? Artificial Intelligence for Ethical Decision Making

---

Dissertation submitted to the
Faculty of Business, Economics and Informatics
of the University of Zurich

to obtain the degree of
Doktor der Wissenschaften, Dr. Sc.
(corresponds to Doctor of Science, PhD)

presented by
Suzanne Tolmeijer
from Arnhem, The Netherlands

approved in February 2022

at the request of
Prof. Dr. Abraham Bernstein
Prof. Dr. Iyad Rahwan

The Faculty of Business, Economics and Informatics of the University of Zurich hereby authorizes the printing of this dissertation, without indicating an opinion of the views expressed in the work.

Zurich, February 16, 2022

The Chairman of the Doctoral Board: Prof. Dr. Thomas Fritz

Abstract

With the advancement of AI technology, an increasing amount of AI applications are being developed and applied in various domains. While some tasks in such applications lend itself well for the strengths of AI, other tasks are more challenging to automate. One example of this is ethical decision making. AI for ethical decision making has not been explored much, among other reasons, for its possibly impactful and ethically loaded results, as well as a lack of 'ground truth' on what is considered the right thing to do. However, AI for ethical decision making could both be valuable in explicit ethical decision making domains and increase the ethical use of other AI applications. This thesis fills the mentioned research gap by focusing on if and how AI for ethical decision making can be designed in a way that is acceptable for users.

The investigated research topics that are part of AI for ethical decision making are presented according to the incremental and iterative design cycle (IID), which is often applied in the development of new technology. After an initial planning phase, a design cycle consists of the following phases: planning and requirements, analysis and implementation, testing, and evaluation.

During the first phase, *initial planning*, we investigate the state of the art of implementing ethical theory in AI, by performing an extensive literature review. Among other results, we find that the field is scattered in terms of the ethical theory and AI types used to create AI for ethical decision making. Additionally, the developed applications consist mostly of prototypes. These results imply that a Wizard of Oz approach is appropriate for the implementation and testing in the design cycle presented in this thesis.

The success of any AI application depends on whether the users trust the AI enough to rely on it. Given the varying opinions regarding a ground truth for ethical AI, where AI decisions can easily be considered to be wrong, we focus on how AI mistakes influence user trust. In the second phase of the design cycle, called *planning and requirements*, we perform an experiment to investigate the effect of AI mistakes and their timing on user trust and reliance. We find that system inaccuracy negatively influences trust and reliance. Furthermore, the negative effect of AI mistakes is stronger when mistakes are made during the first interaction with the user.

To mitigate these negative effect of AI mistakes, the third phase of *analysis and implementation* focuses on AI mistakes and how their negative effects can be mitigated, by presenting different interaction designs. This is done by introducing a taxonomy of AI mistakes and appropriate mitigation strategies.

In the fourth *testing* phase, we use a Wizard of Oz application to test user perception of AI for ethical decision making. We find that while participants had higher moral trust in a human expert and find humans more responsible, they had more capacity trust and overall trust in an AI system for ethical decision making.

In the final phase, *evaluation*, we describe the consequences of our finding. Since people perceive AI and humans to have different strengths that are both valuable for ethical decision making, we propose an interaction paradigm that utilizes the strengths of both: human-autonomy teaming. For AI and humans to be able to form an effective team, further development of different AI capabilities is needed: agency, communication, shared mental models, intent, and interdependence.

In conclusion, this work contributes to the understanding of user perception of AI for ethical decision making, and suggests design strategies to move research on AI for ethical decision making forward.

Acknowledgements

Life is like a roller coaster. So is doing a PhD. Handling both at the same time while being a sensitive person can be.. well, like a roller coaster on fire without properly working breaks. Sometimes, you close your eyes and hold on in shear terror, just hoping for the ride to end well. Other times, you throw your hands up in the air and enjoy the ride. Let's just say, it's been intense, and I am thrilled that I made it to the point where I can write this acknowledgements section. Especially since there are a lot of people to acknowledge — and I am going to take my time and bask in this moment, mind you.

Firstly, I want to thank my prof and advisor Avi. Besides being very good at what he does (and he does A LOT), he is exceptionally good with people. I don't think I could have managed this ride without his support, during the good, but especially during the challenging times — how's that tissue stash in your office doing? I have learned so much from you, which I hope to be able to apply and pass forward to others during the rest of my career. Also, and blissfully so, 'The Bern' has been etched into my memory for eternity.

Secondly, I want to thank our excellent postdocs. Whenever Avi was too busy doing the many things he does, I could always count on them for academic and personal support. Cris has seen and helped me grow from a insecure student who had no idea what she wanted to do, let alone how, into a full-fledged researcher who's ready to take on the academic world. While Luca joined our group at a later stage, he has been equally supportive, taking on yet another project or (feedback) task to help while already having plenty on his plate. It has been noticed and very much appreciated.

Next, I want to thank all my present and past colleagues at DDIS. It was a delight to experience the different interactions I had with them, from questioning junior with Bibek, Tobi and Daniel, to a struggling equal with Romi, Narges, Matthias and Martin, to an outspoken senior with Kat, Flo, Lucien and Rosni. I will miss Karl Wiesenknecht the most though.

In addition, many thanks to all my collaborators, who have challenged me and helped me in many ways. Doing multi-disciplinary collabs are very challenging, but end up helping you grow as a researcher and as a person. Thanks to all paper reviewers who improved the papers presented in this thesis, and to Iyad for agreeing to be the external reviewer for this thesis. Thanks to all research friends I made that I could brainstorm with — I still hope to work together with some of you if we haven't yet. A special thanks to Stefan Schlobach, my former professor from Amsterdam. He did not only got me back to human-centered computing when computer-centered computing didn't cut it for me, but showed me the opportunities of science and introduced me to Avi.

Privately, I want to thank my old friends who stuck by me despite the distance to the Netherlands, who warmed my heart whenever it needed warming. You know who you

are. Equally important, I want to thank all the new friends I made during my time in Switzerland that made it feel like home. Both in and outside of IfI, I feel like I have made friends for life.

Finally, I want to thank my family for always being there (especially Larry and Larry), and my bearded hero, for being a shoulder to cry on whenever this strong independent women needed a break from the world.

Thanks to everyone who helped me finish this ride in style. Onwards to the next!

Suzanne
Arnhem, November 2021

Financing
This work is partially funded by armasuisse Science and Technology (S+T), via the Swiss-Center for Drones and Robotics of the Department of Defense, Civil Protection and Sport(DDPS).

Table of Contents

I Synopsis

1 Introduction . . . 3
1.1 Motivation . . . 3
1.2 Approach . . . 5
1.3 Background and Research Questions . . . 6
1.4 Contributions . . . 14
1.5 Outline and Contribution Statements . . . 19

II Contributions to Thesis

Implementing Ethics in AI

2 Implementations in Machine Ethics: A Survey . . . 24
Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein
2.1 Introduction . . . 24
2.2 Introduction to Machine Ethics . . . 26
2.3 Survey Methodology . . . 28
2.4 Object of Implementation: Ethical theories . . . 31
2.5 Non-Technical Implementation Aspects . . . 36
2.6 Technical Implementation Aspects . . . 41
2.7 Analysis . . . 46
2.8 Future avenues and limitations . . . 53
2.9 Conclusion . . . 56
2.10 Acknowledgement . . . 56

AI Mistakes and Trust Formation

3 Second Chance for a First Impression? Trust Development in Intelligent System Interaction . . . 58
Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein
3.1 Introduction . . . 58
3.2 Related work . . . 60
3.3 Study Design . . . 62
3.4 Results . . . 66

3.5 Discussion And Future Work 73
3.6 Conclusion 75

4 Taxonomy of Trust-Relevant Failures and Mitigation Strategies 77
Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman
4.1 Introduction 77
4.2 Related work 78
4.3 Proposal 82
4.4 Autonomous Trust Repair 88
4.5 Future Work 92
4.6 Conclusion 93

AI for Ethical Decision Making

5 Capable but Amoral? Comparing AI and Human Team Members in Ethical Decision Making 95
Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein
5.1 Introduction 95
5.2 Related Work 98
5.3 Method 101
5.4 Results 107
5.5 Discussion 112
5.6 Conclusion 115

Human-Autonomy Teaming

6 Human-AI Teaming in the Cockpit: Domain Mapping and Research Agenda 117
Suzanne Tolmeijer, Fabio Mattioli, Simon Coghlan, Martin Tomko, and Natasha Sutila
6.1 Introduction 117
6.2 HAT Theory for the Aviation Domain 118
6.3 Research Agenda for HAT in Aviation 128
6.4 Conclusion 136

III Conclusions

7 Conclusions 141
7.1 Limitations and Future Work 141
7.2 Conclusion 143

IV Appendix

A Descriptions of Selected Papers in Chapter 2 147

List of Figures 156

List of Tables 157

References 159

Curriculum Vitae 197

Part I

Synopsis

Chapter 1

Introduction

1.1 Motivation

Since its birth in the 1950's, artificial intelligence (AI) has been growing rapidly as a research field. Using different approaches, such as logic-based expert systems during the 70's and 80's (e.g., [41, 224, 459, 508]) and unsupervised machine learning since the 2000's (e.g., [370, 372, 517, 528]), the field has attempted to "*not just understand but also to build intelligent entities*" [421, p 1]. Given the breadth of this mission, the term AI has been used as an umbrella term, encompassing many different 'intelligent' algorithms[1]. As such, it has been difficult to define the term in a way that is both specific enough to carry value, but broad enough to include all approaches used in the field. Russell and Norvig [421] define it as follows: AI can be considered as an agent that is "perceiving its environment through sensors and acting upon that environment through actuators" [421, p 34].

There have been many successful applications of AI algorithms that influence people in their everyday lives. Some software examples are health care diagnostics [258] and drug discovery [94], automated financial investments [450], recommender systems that learn your preference to recommend products or services [79], and marketing chatbots that represent to answer client's questions [312]. Following the popularity of smartphones, many people carry AI applications with them everyday in their pockets, such as voice assistants [25] and smart map planning [434]. Hardware applications have also increased, such as autonomous factory robots working assembly lines [428], consumer robots that vacuum your house [27], and cars that drive themselves [407].

Many of these applications have provided new services and products to become available to the general public, or have improved existing ones, e.g., by increasing speed or accuracy. However, there have also been unintended negative side-effects of using AI in practice. For example, contextual bias can result in health care AI generalizing across patient groups or service settings (such as rural areas in a third world country versus a high-end hospital in the first world) when this is inappropriate [400]. Unbalanced training data can lead to biased algorithm results, such as racial bias in face recognition software due to a lack of Afro-American and Asian examples [88]. Existing biases run the risk of being perpetuated and magnified, such as when hiring algorithms show a gender preference because current employees have an unbalanced worker population [539]. A classification algorithm might predict high confidence on an error because the class was an 'unknown

[1] Because of the breadth of the term 'AI', the following terms are used as additional synonyms throughout this thesis: algorithm, system, and machine.

unknown' not present in the training set [32]. In sum, there are many reasons that can cause AI to make mistakes or produce biased results. Some of these issues might only have small effects, such as a recommender algorithm recommending an item you are not interested in. Others, however, can have a more grave impact. When AI systems are used without human supervision, people might not be hired [539], might not get parole [257], be rejected for a loan [118], or not get diagnosed correctly [14] because of algorithmic error and bias.

One of the AI applications that has been explored less is AI for ethical decision making. Example applications could be related to biomedical dilemmas such triage decisions, sentencing in law cases, or risk management in national defence scenarios. One of the reasons these applications have been explored less is that the possible negative consequences of AI for ethical decision making can be far reaching, to the point that they can result in life-and-death decisions. Another issue is that of responsibility: in case of negative consequences, AI cannot be held accountable in a court of law as of yet. Moreover, algorithmic errors imply there is a correct answer that the AI failed to produce. However, philosophers have been discussing and disagreeing for centuries on a unifying ethical theory. If there is no agreed-upon ground truth, it is unclear what the AI should learn to reproduce. Finally, public perception of AI influences adoption of certain applications.

Discussions on ethical decision making for autonomous cars [525] and autonomous weapons systems [475] have shown that i) public discussions on ethics of AI are not always related to actual technical challenges but to the general public's perception of those challenges, and ii) general impressions of AI shape these perceptions, which are influenced by news and other media [162]. Science fiction stories in particular have amplified hopes and fears related to AI, such as killer robots dominating the human race or the future scenario of stress-free lives because AI takes over all jobs [89].

Current research on AI for ethical decision making has focused on prototypical applications of ethical theory in AI and general perception of different sorts of AI. Nevertheless, what has been missing is a more thorough research on people's perceptions of AI for ethical decision making, and the implications for its potential. To fill this gap, this thesis focuses on different aspects that influence perception of AI for ethical decision making:

- How far advanced is the current technology?
- How do perceived mistakes influence AI perception (which is especially relevant in ethical decision making with severe consequences)?
- How can AI be designed to mitigate negative influences of AI errors?
- How do people perceive AI versus a human for ethical decision making?
- What possible application forms could add value in practice?

To summarize, this thesis considers the following research question:

Can an AI application for ethical decision making be designed in a way that is acceptable for users, and if so, how can it be applied?

1.2 Approach

To explore this question, this thesis will make use of the incremental and iterative design (IID) process often used for developing new software applications and HCI designs (see Fig. 1.1). It works as follows: after initial planning and research on the current state of the art, this approach follows an iterative and incremental approach to design a new system. After a planning phase where requirements are gathered, the next phase consists of analysis of the requirements, as well as the design and implementation of the system. The system is tested, and the evaluation informs a next and improved iteration of the system. Once the system meets the set requirements, it can be deployed.

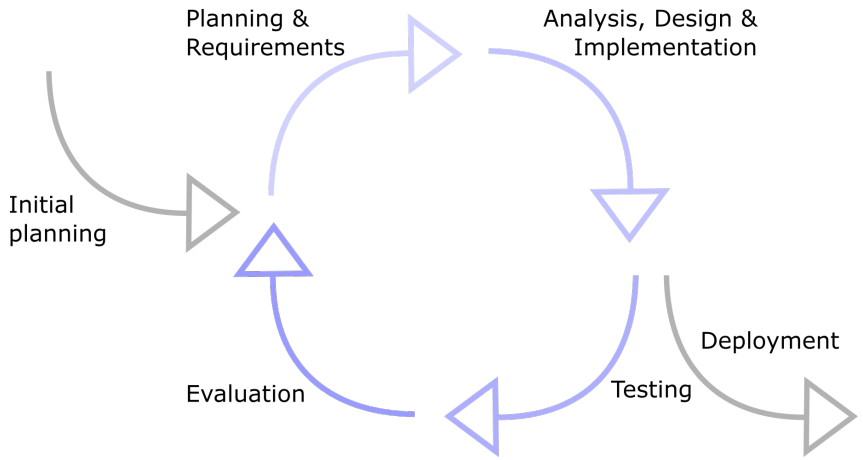


Fig. 1.1: Iterative and incremental design (IID) cycle, adapted from Basil and Turner [45]. See [290] for the history of the model.

In the developmental process, what the implementation of a system looks like highly depends on the current technical state of the art. When the technology is readily available, it is possible to try out different version of the system. However, when the technology used is still being developed, it is possible to use a Wizard of Oz (WOz) approach to envision what a finished system could look like. This approach is applied in "*studies where subjects are told that they are interacting with a computer system through a natural-language interface, though in fact they are not*" [122, p 194]. As can be seen in Figure 1.2, the WOz approach can be used to bridge the gap between the technical state of the art and envisioned system, in order to i) get requirements for the intended end system and ii) steer the development process based on intermediate results. Depending on the analysis of the state of the art in the first part of this thesis, the appropriate amount of WOz can be applied in the second part of this thesis.

The next section of this thesis introduces the relevant background of each part of the thesis, leading to the research questions that are investigated.

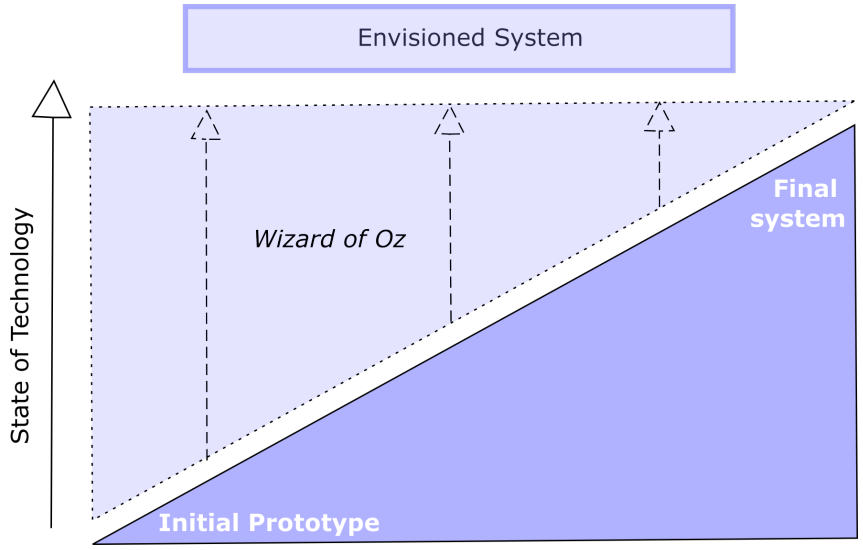


Fig. 1.2: Usage of Wizard of Oz at various stages of system design, adapted from Dow et al. [146].

1.3 Background and Research Questions

By applying the IID model, this section is divided into different steps. For each part of the design process, related work is presented leading to the research question for that design phase.

1.3.1 Initial Planning Phase — Where are we now?

To understand the starting point of designing AI for ethical decision making, an assessment is needed of the current state of the field. In his influential paper, Moor [348] argues that by nature, computing technology is normative, since the intended purpose serves as a norm for evaluation. He presents four different types of ethical agents: ethical-impact agents, implicit ethical agents, explicit ethical agents, and full ethics agents. Ethical-impact agents do not perform any ethical act, but simply because they exist, they have ethical consequences (such as taking the job of a human). Implicit ethical agents do not explicitly focus on ethics, but are designed in a way that promotes ethical outcomes. Examples of these are systems designed for safety and reliability. Explicit ethical agents have ethical theory, such as consequentialism, explicitly implemented in their decision process. Finally, a full ethical agent is an agent that "*can make explicit ethical judgments and generally is competent to reasonably justify them*" [348, p 20]. At the time of writing, he only considered humans to be full ethical agents, and questioned if and how AI could become a full ethical agent.

There are some issues with this classification: these definitions do not allow for a technical distinction between AI applications, and on a philosophical level, Moor never defines what a moral agent is. Nevertheless, this classification still gives a first indication of the levels of ethics that could be embedded in AI. While the first two types have existed for a long time and the latter is currently out of reach, the potential of the explicit ethical agent is worth exploring.

The field that explores implementing ethical theory in AI, also called '*machine ethics*' [11], appears quite scattered. Researchers focus on different types of tasks, implement different types of ethical theories, while using different types of AI technology. Examples range from authors using logic-based AI to represent deontological ethics [74] to authors using machine learning to have the AI perform consequentialist actions [2]. Given the seemingly scattered nature of the field, a structured literature review is needed to understand the state, gaps, and potential of the field. This leads to the first research question:

Research Question 1: What is the state of the art of implementing ethical theory into AI?

Implementing ethical theory into AI has an ethical component (i.e., which theory is being implemented), a technical component (i.e., which computational approach is used to implement the theory), and an implementation component (i.e., how is the theory implemented in practice). For each component, we expect the following, respectively:

Hypothesis 1: Most AI applications that implement an ethical theory use deontological ethics.

Broadly speaking, there are three main streams of objective ethical theory: *deontological ethics* (e.g., [6]), focusing on intention during a decision, *consequentialism* (e.g., [461]), focusing on outcome of a decision, and *virtue ethics* (e.g., [259]), focusing on the general character of the decider. Since AI does not have a general character to develop virtues and consequences of actions can be so far-reaching that they become incalculable, we hypothesize that deontological ethics are applied most in AI implementations.

Hypothesis 2: Implementations follow a top-down approach and mostly focus on action selection.

Implementations of ethics can be done in two ways. The first is a top-down approach, where existing theory and knowledge is translated into implementable code. The second is a bottom-up approach, where ethical knowledge is learned or derived from data. We expect the top-down approach to be used more, since i) there is much ethical theory and knowledge that could be utilized, and ii) there is no guarantee that a learning AI will learn the preferred behavior for each possible case [256, 385]. Additionally, we hypothesize that authors will focus most on one ethical approach to select an action for the AI, rather than focusing on ethical model selection in an intermediate step.

Hypothesis 3: Logical reasoning is applied more than learning algorithms.

Recently, machine learning has seen a massive increase of application. However, following Hypothesis 2, we hypothesize that existing ethical knowledge will be implemented using logical reasoning over learning techniques, since a reasoning algorithm lends itself better for formalizing and applying existing knowledge in an algorithm.

1.3.2 Planning and Requirements Phase — What do we need?

Part of the sensitivity of ethical decision making comes from the fact that 'mistakes' can lead to severely negative and ethically loaded results. For people to accept and use AI for ethical decision making, they need to trust the system to advice or do the right thing without perceived mistakes.

There are many different factors that influence trust in AI. Hoff and Bashir [233] have summarized empirical research on trust in AI in an overarching framework.[2] According to them, the different factors that influence trust can be divided into dispositional trust, situational trust, and learned trust. Dispositional trust is influence by factors like culture, age, and gender, and is considered to be a user's general tendency to trust, independent of the system or situation. Situational trust, on the other hand, depends on the situation in which the system is deployed. There is both internal variability in the situation which influences trust, such as self-confidence and mood, but also external variability, such as the workload and framing of the task. Finally, learned trust consists of initial learned trust, based on pre-existing knowledge, and dynamic learned trust, which develops during interaction with the system. Especially during dynamically learned trust, system errors can influence trust formation of the user.

There has been some research on the factors that affect how an AI's mistakes influence trust formation. For example, anthropomorphizing the AI has an influence [487], the age of the user experiencing the error [231], and the domain expertise of the user [366]. However, there has been little research on trust formation during various interactions over time. This is surprising, since technology use in practice consists of exactly that: multiple interactions over a longer span of time. The research that comes closest has all focused on trust formation over time within one single user session. For example, Holliday et al. [237] looked at the influence of explanations on trust formation within one user session: they found that explanations temporarily increase user trust, but found that perceived system ability has a larger influence. Consistent reliability steadily increases trust, while consist unreliability decreases trust [50]. Nevertheless, the combination of system mistakes and trust formation during multiple user sessions has been under-researched. AI mistakes, such as perceived unethical AI actions in the context of AI for ethical decision making,

[2] While the authors talk about 'automation' rather than AI, their definition has high overlap with the definition of AI used in this thesis: "*technology that actively selects data, transforms information, makes decisions, or controls processes*" [233, p 408].

can happen during any of those interaction. This leads us to the second research question:

Research Question 2: How do mistakes of an AI influence trust formation over time?

Mistakes are expected to both influence trust formation and reliance on the AI system. As such, we hypothesize the following:

Hypothesis 4: Accurate advice leads to user reliance.

We hypothesize that when an AI system gives accurate advice, the user is more likely to (continue to) rely on the system [541]. At the same time, inaccurate advice will lead to (continued) lack of reliance on the system, where the user will try to manually complete the task themselves [142].

Hypothesis 5: Inconsistent accuracy of advice leads to lower trust.

We expect that trust is directly influenced by how consistent the AI system is in its advice. In case of inconsistent advice, trust is expected to decrease, while consistent advice is hypothesized to increase reported trust [441].

Hypothesis 6: Timing of inaccurate advice influences trust formation: earlier mistakes have a higher impact on trust formation.

The trust users place in an AI system is influenced by the expectations they have of the system's performance. When users have less experience with a system, they will base their trust on the first experiences they have. Hence, we hypothesize that mistakes made earlier on in the system usages will have a more negative impact on user trust then mistakes made later [141].

1.3.3 Analysis and Design Phase — What should it look like?

Because AI mistakes can influence a user's trust levels, it is beneficial to take this into account when designing AI for ethical decision making. It would be undesirable to lose user trust because of an error unrelated to its ethical decision making capacities. One way to take this into account during AI design, is by using trust recovery strategies. Compared to *trust formation* in AI, *trust loss* and *recovery* have been investigated less [130, 286]. Robinette et al. [416] showed that timing and exact content of trust repair messages influence the effectiveness of the intervention. Additionally, Kohn et al. [286] found that a high human-likeness of AI leads to higher perceived intent, which in turn leads to more severe trust loss when an error occurs: it becomes harder to gain trust back with trust recovery strategies such as an apology. This was confirmed by Kim and Song

[282]: they found that anthropomorphism influences which type of trust repair strategy is more effective. The effect of anthropomorphism can be explained by the fact that the concept of trust repair, both theoretically and empirically, stems from human-human interaction [130]. However, people do not trust humans and AI in the same way. In the case of robots, overall trust was similar to trust in humans, but mistakes had a larger negative influence on trust in the robot's case [483]. This effect has been replicated for many different domains and different forms of AI [229].

While empirical work has been done to research specific AI mistakes and trust recovery strategies, an overview has been missing that aggregates the different trust loss and mitigating actions for AI applications. Since AI can make different errors from humans, such as having issues because of a faulty software update, an overview is needed of specific AI mistakes that cause trust loss. Moreover, for design purposes, it would be useful to have an overview of fitting mitigation strategies for each type of AI error, to attempt trust recovery. This leads us to the following research question:

Research Question 3: Which errors of AI can lead to trust loss and how can we mitigate the consequences of errors?

We expect that different mistakes and different mitigation strategies can be relevant, which are represented in the following hypotheses.

Hypothesis 7: Different types of AI mistakes that have an effect on trust formation can be distinguished.

We hypothesize that mistakes can be caused by both the AI and the user, which both influence trust. Additionally, one can differentiate between mistakes that are not intentional, or intentional system behavior that is perceived by the user as faulty [345].

Hypothesis 8: Effective types of mitigation strategies to recover trust loss depend on the type of AI mistake that caused the trust loss.

We expect that not all mitigation strategies are appropriate for all types of AI errors. For example, we expect that providing the user with an alternative solution to their request is only relevant when a system error causes their request to not be possible [416].

1.3.4 Testing Phase — What do people think?

In the next step of this design cycle, the focus is on testing AI for ethical decision making. The (Wizard of Oz version of an) AI is presented to users, to evaluate user perception and design requirements.

Since AI is increasingly taking over tasks that were initially executed by humans, one way of testing user perception of AI is by comparing perception with a human performing an equivalent job. Research thus far has shown that people judge machines more by the outcome of their actions, while humans are judged by their intentions [229]. Additionally, they judge machines more harshly in the case of negative outcomes [229, 243]. In general, two categories of AI perception tendencies can be distinguished: *algorithmic aversion* [142] and *algorithmic appreciation* [310]. Algorithmic aversion mostly stems from the AI making mistakes [142], and leads users to lose trust. Specifically, the expectations users have, timing of mistakes, and consequences of mistakes influence algorithmic aversion [293]. Algorithmic appreciation, on the other hand, can stem from the 'machine heuristic': people expect a machine to be more objective and less biased than a human [474], and therefore trust the AI more than a human. In an attempt to reconcile the two concepts, recent work found that how the (expertise of the) human and AI are framed, as well as domain expertise of the user, influence whether algorithmic appreciation or aversion is triggered [252].

In the context of AI for ethical decision making, trust and distrust have not been investigated much, possibly since there are not many AI applications in everyday life yet that make ethical decisions. The most investigated domain is the autonomous cars domain. A literature review from Feroz et al. [166] summarized that while people have mixed opinions on autonomous cars, they worry most about the legal and ethical implications of such machines. Their trust on the other hand is mostly related to safety concerns. The fact that people have different preferences with regards to the ethical guidelines of autonomous vehicles has been shown on a large scale [34]. However, the question remains open if people's ethical preferences and consequential trust in AI generalize beyond autonomous cars, or depend on the domain in which ethical AI is applied. To this end, a more thorough investigation of a different domain than autonomous cars for AI ethics would further our understanding of people's perceptions of ethical AI.

Additionally, when considering the perception of AI making ethical decisions, it is relevant to investigate the responsibility people assign to the AI. Especially in the context of ethical decision making, decisions can have large and ethically charged consequences, which need to be accounted for. From a philosophical perspective, it has been argued that for an entity to be held morally responsible, it needs to be a moral agent — the requirements of which no machine has achieved as of yet [378]. Instead, responsibility and the following liability could be distributed between all involved parties, such as designers, regulators, and users [478]. This approach can explain the explosion of AI ethics guidelines that have been published in recent years [270], where many involved parties are suggested to be involved in the creation and maintenance of ethical AI. However, with the rise and success of machine learning, the fear has arisen that a responsibility gap will occur: the creator is not able to predict the exact behavior of the AI, therefore not being able to be held responsible, while the AI can also not be held responsible since it does not have

(moral) agency [331]. Not everyone agrees with this premise in theory (e.g.,[486]) and first indications contradict the premise in practice. In bail decisions, causal responsibility and blame were assigned similarly to humans and AI, although other types of responsibility (such as responsibility-as-obligation) were assigned more to human agents [305]. A possible explanation for the assignment of blame and moral responsibility to AI is that people assigned mental states to AI similarly as to humans [472]. As with trust in AI for ethical decision making, it would be beneficial to research responsibility assignment in a different domain from the 'standard' autonomous vehicle domain.

To investigate both trust and responsibility perceptions for ethical decision making, an AI for ethical decision making was developed in collaboration with a game development company.[3] The results from Research Question 1 will determine the level of WOz needed for the application. Using this application, we can investigate the following research question:

Research Question 4: How do people perceive AI for ethical decision making?

Until recently, ethical decision making has mostly been a human endeavor. As a consequence, we expect people to prefer humans over AI for ethical decision making. This is formalized in the following hypotheses:

Hypothesis 9: People trust a human more than an AI for ethical decision making.

As mentioned, people display both algorithmic appreciation and algorithmic aversion towards AI [252]. Given the sensitive nature of ethical decision making, we hypothesize that algorithmic aversion will predominate people's perception of AI for ethical decision making. Instead, we expect that human decision makers are trusted more.

Hypothesis 10: People perceive an AI for ethical decision making to be less responsible than a human making ethical decisions.

Earlier studies have shown a responsibility gap when AI was making decisions [323, 472]. While this gap has not always been as apparent in practice as hypothesized in theory [331], we still hypothesize that participants will hold humans more responsible for ethical decision making that AI.

Hypothesis 11: People rely less on AI for ethical decision making than a human equivalent.

Following Hypothesis 9 and Hypothesis 10, we expect that participants will rely less on AI for ethical decision making than on humans, since we expect that trust in AI [95] and responsibility assignment to AI [298] influences reliance.

[3] Koboldgames: https://www.koboldgames.ch/

1.3.5 Evaluation Phase — What's next?

The results from Research Question 4 show that people perceive humans and AI to have different capabilities, in which they are correct: machines currently outperform humans in many tasks, such as pattern recognition and optimization, while humans still outperform AI in common sense, communication, and explanation [5]. A possible way to utilize the strengths of both parties, is by using human-AI teaming (HAT) as an application form.

Thus far, technology has been utilized as a tool — a means to an end in a digital rather than analogue form. However, the combination of increased AI capabilities, potential for autonomous application, and natural language interactions have started the debate on whether AI can be more than a tool [316].

Some discard this notion completely. In her provocative article 'Robots should be slaves', Bryson [78], argues why we should not see AI at more than tools. She recommends that AI should not be designed to have anthropomorphic features, to not give the impression of being more than a tool. Others argue AI can be more than a tool, either because it actually has the capabilities [136] or because users perceive the system as such [546]. Yet, for AI to be perceived as a potential team mate, it needs to fulfill various requirements. Lyons et al. [316] summarize these as follows: AI needs to have agency, communication with human team members, a have a shared mental model, have intent towards a shared goal, and have interdependence with its human team members. Since AI will not completely take over all tasks right away, it is important to focus on the intermediate phase of human-AI collaboration in a teaming setting [272].

Human-AI teaming is a fairly new concept and has been investigated mostly in a theoretical manner (e.g., [93, 148, 272, 337, 457]). Empirical work so far has shown there are still many open questions regarding human perception and AI capabilities [371]. Because of the theoretical nature of most work on the topic, a gap exists for AI researchers on what HAT could like like in specific application domains. To make HAT more tangible while investigating its potential, we present the domain of aviation as an investigative tool. It makes for an excellent example domain, since aviation work is highly protocolized, involves extensive skills and training, multiple team members in the cockpit and on the ground, and takes place in a context that is complex and yet somewhat predictable. Using this example domain, we present the last research question:

Research Question 5: Could Human-AI Teaming be a useful application format for collaborating with AI?

The appropriateness of HAT for ethical decision making depends on the domain in which it is applied. Because the study that is conducted to answer Research Question 4 focuses on the aviation domain, we specifically focus on HAT for aviation and utilize the knowledge gained for this next step. We hypothesize the following:

Hypothesis 12: More sophisticated modes of interaction are needed than the currently limited natural language options of text and voice.

As mentioned before, proper teaming dynamics include a shared understanding of a common goal and continuous communication while working towards this goal. Decision cases that include ethical dilemmas often have many aspects that need to be considered in the context of the decision. We expect that more research is needed on interactions modes that make these detailed considerations possible in a HAT setting, since current empirical work has mostly been done in a WOz setting due to the lack of technical sophistication [316].

Hypothesis 13: The team composition of human-AI teams influences the type of application it could be used for.

We hypothesize that both the number of team members and their respective roles influence if and how HATs can be applied to tasks. In the case of ethical decision making, how responsibility is shared among team members is of particular importance. It has been shown before that when users feel they cannot assign responsibility to AI, they are less likely to use it in decision making settings [402].

1.3.6 Summary of research questions

The combined research questions form one cycle of the IID cycle to research AI for ethical decision making. The summary of the research questions researched in this thesis, plotted on the IID cycle, can be found in Figure 1.3. An overview of the research questions and hypothesis can be found in Table 1.1.

In each of the subsequent design steps, many factors can be looked at for the design of AI for ethical decision making, especially in the requirements analysis and subsequent design phase. The focus on trust in Research Question 2 and Research Question 3 stems from the knowledge that "*Trust between algorithms and human agents is the underlying key factor in the heuristics and acceptance of AI*" [456, p 7]. The chosen topics for each phase were selected because they were deemed most appropriate to answer the main research question: "***Can an AI application for ethical decision making be designed in a way that is acceptable for users, and if so, how can it be applied?***".

1.4 Contributions

The posed research questions are explored in depth in the papers presented in Part II of this thesis. This section briefly summarizes the findings related to each research question.

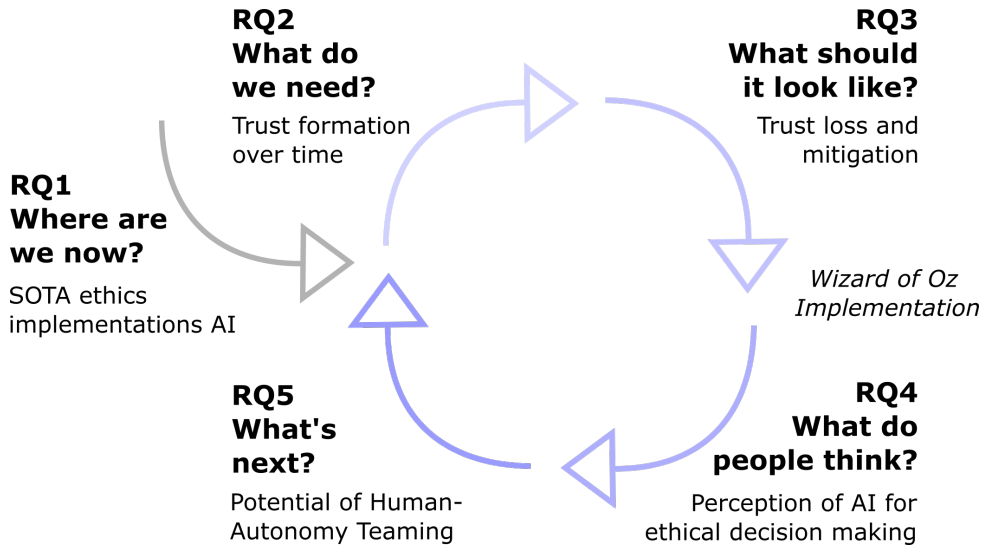


Fig. 1.3: Summary of the research questions presented along the IID cycle.

1.4.1 Implementations in Machine Ethics: A Survey

The first research question regards the state of the art of machine ethics, i.e., implementing ethical theory into AI. To answer Research Question 1, we conducted a literature survey to summarize the status of the field, presented in detail in Chapter 2.

Given that machine ethics is a multidisciplinary field that combines philosophy and computer science, this literature review was conducted in collaboration with moral philosophers. After calibrating our understanding of the relevant concepts, we performed an extensive literature review, focusing on the keyword 'implementation' combined with various synonyms of machine ethics. We introduce three taxonomies according to which the found papers were classified: ethical theory, non-technical implementation aspects, and technical implementation aspects.

In terms of ethical theory, we found that most implementations use a single ethical theory in their implementation, which are mostly deontological ethics or consequentialism. This partially confirms Hypothesis 1. Findings for the non-technical aspects of implementation include that 1) most implementations are done in a top-down fashion, 2) most do not focus on a specific domain, 3) about half of the implementations focus on action selection and execution, and 4) approximately half of the implementations do not perform a formal evaluation of their system (and seem to be in a prototype stage of development). These findings confirm Hypothesis 2. Regarding the technical aspects of implementation, we find that most authors use logical reasoning for their systems or a hybrid of multiple algorithmic approaches, which confirms Hypothesis 3. However, very few provide the code base or details of their implementation.

Table 1.1: overview of research questions and hypotheses covered in this thesis.

RQ1	**What is the state of the art of implementing ethical theory into AI?**
H1	Most AI applications that implement an ethical theory use deonto-logical ethics.
H2	Implementations follow a top-down approach and mostly focus on action selection.
H3	Logical reasoning is applied more than learning algorithms
RQ2	**How do mistakes of the AI influence trust formation over time?**
H4	Accurate advice leads to user reliance.
H5	Inconsistent accuracy of advice leads to lower trust.
H6	Timing of inaccurate advice influences trust formation: earlier mistakes have a higher impact on trust formation.
RQ3	**Which errors of AI can lead to trust loss and how can we mitigate the consequences of errors?**
H7	Different types of AI mistakes that have an effect on trust formation can be distinguished.
H8	Effective types of mitigation strategies to recover trust loss depend on the type of AI mistake that caused the trust loss.
RQ4	**How do people perceive AI for ethical decision making?**
H9	People trust a human more than an AI for ethical decision making.
H10	People perceive an AI for ethical decision making to be less responsible than a human making ethical decisions.
H11	People rely less on AI for ethical decision making than a human equivalent.
RQ5	**Could Human-AI Teaming be a useful application format for collaborating with AI?**
H12	More sophisticated modes of interaction are needed than are currently available.
H13	The team composition of human-AI teams influences the type of application it could be used for.

The status of the field can be considered to be a 'polycentric oligarchy': several independent groups of researchers confirm each other's assumptions and do not communicate much with other clusters that hold different views. Lack of standardization and formalization poses a severe problem to the development of the field.

Moving forward, it would be beneficial to combine multiple ethical theories, focus on domain-specific ethics with higher consensus among domain experts, and include folk morality to increases chances of acceptance and adoption in society. Additionally, more systematic evaluation is necessary, which could be facilitated with domain-specific benchmarks. Collaboration between different research fields could be beneficial to make practical applications a reality. Finally, machine ethics development should focus on transparency of the system, share code bases for further development, and consider the usage of feedback by users.

1.4.2 Trust Development in Intelligent System Interaction

Research Question 2 and Research Question 3 focus on the impact of errors made by the AI on trust formation of the user. Chapter 3 provides the answer to Research Question 2 by describing an investigation into the influence of AI mistakes over time on trust formation. Specifically, we investigated if the accuracy of the system influences reliance on the system, whether and how (in)consistency of accurate advice influence trust formation, and if dispositional factors such as age, gender, and affinity with technology influence user trust.

To do so, we developed a web app where crowdworkers were asked to fulfill housing search tasks with the help of an intelligent system. They were asked to participate in three separate sessions with two days in between, to understand how trust in the system formed over time. They could choose whether or not to use the system during their decision process. Depending on the experimental group, the system would give correct or incorrect advice in the first, second and/or third session.

We found that system inaccuracy lowered reliance on and trust in the system, confirming Hypothesis 4 and Hypothesis 5. Additionally, first impressions had a large influence on trust formation: the average trust score for one initial inaccurate advice was the same as users that got a correct first advice followed by two sessions of incorrect advice. This confirms Hypothesis 6. Trust recovery is possible, which seems to be explained by the fact that the system was perceived to be learning over time.

In sum, accuracy of the system and timing of mistakes have a large influence on user reliance and trust formation.

1.4.3 Taxonomy for Trust-Relevant Failures and Mitigation Strategies

Given that mistakes have such a large impact on trust, system design should include mitigation strategies for system inaccuracy when appropriate. To answer Research Question 3, Chapter 4 presents an overview of which failures can influence trust formation and how negative effects can be treated.

By investigating related work and experiences from trust research 'in the wild', we found four distinct failure types that can influence trust formation: system failures (both hardware and software), design failures, expectation failures (either expecting something that does not happen or not expecting something that happens), and user failures (which can be intentional or by accident). This confirms Hypothesis 7.

Possible mitigation strategies, which were mapped to the found failure types, include improved interaction design, providing explanations to the user, apologizing for mistakes made, fixing any system issues that occur, proposing an alternative, and providing user training. This supports Hypothesis 8.

1.4.4 Comparing AI and Human Team Members in Ethical Decision Making

The next step in the design of AI for ethical decision making included the implementation of a system and testing of user perceptions of such a system, presented in Chapter 5. The results of Research Question 4 showed that the field is in the phase of initial prototypes. As such and following Figure 1.2 presented in Section 1.2, it is warranted to use a WOz approach to research system requirements for AI making ethical decisions in this initial stage of system design. Additionally, the results of Research Question 2 indicated that mistakes and timing of mistakes have a massive effect on user trust and perception. However, given the fact that basic understanding of the perception of AI for ethical decision making still needs to be explored in this domain, we chose to exclude obvious system errors in this first exploration of user perception, since they could confound this first insight into user perception.

In collaboration with game development company Koboldgames, we developed a WOz prototype of an AI making ethical decisions. We chose the domains of Search and Rescue (SAR) as well as the defense domain, as these are two domains where autonomous aviation systems can be imagined in the near future (e.g., [3, 70, 143]). To analyze the perception of participants, we focused on three dependent variables of perception: trust, reliance, and perceived responsibility. We created two scenarios: participants had to focus on maximizing the amount of lives saved in a SAR setting or minimizing the amount of lives lost in a defense setting. They were presented with both a human expert and AI expert, that would either give them advice or decide what would happen. In the former case, participants had to decide what would happen, in the latter they could veto the decision and choose another option if they disagreed with the choice.

We found that participants had higher moral trust in the human expert, but had a higher capacity trust and overall trust in the AI, which partially confirms Hypothesis 9. These trust beliefs also showed in user reliance: towards the final missions, they relied more on the AI expert than the human expert. These findings on trust and reliance imply we have to reject Hypothesis 11. However, they deemed the human expert significantly more responsible than the AI expert. Instead, programmers and sellers of the AI were deemed partially responsible. This confirms Hypothesis 10.

People's perception appear to be in line with the discussion on meaningful human control [430]: while AI might have certain capabilities in which they excel humans, humans are more trusted for the moral aspect of decision making and deemed more responsible for the decisions that are made.

1.4.5 Human-AI Teaming in the Cockpit: Domain Mapping and Research Agenda

In the final step of the incremental and iterative design cycle, the results of Research Question 4 need to be evaluated to determine which requirements can be used for a next

iteration of the design. The results indicated that while AI is perceived to have capabilities that can add to the ethical decision making process, humans excel at ethical decision making and need to be involved in the final decision. One possible solution, which answers Research Question 4, can be the use of human-AI teaming, which is further examined in Chapter 5.

Chapter 5 summarizes the different components that are needed for HAT: agency, communication, shared mental models, intent, and interdependence. The system should be designed for appropriate reliance and can deploy various kinds of interaction modes. To become a full teammate, it needs emotional intelligence on top of task specific knowledge. HAT research can take inspiration from human-human teaming, as well as from human-animal teaming.

Future research in HAT can look at team composition and level of autonomy of the team members, modes of interaction, how emotional intelligence can be developed and displayed, and focus on the ethical consequences of implementing HATs in practice. The results of this chapter confirm Hypothesis 12 and Hypothesis 13.

1.5 Outline and Contribution Statements

The remainder of this thesis consists of the chapters listed below. Given that all papers featured in this theses had multiple authors, the main author's contribution per chapter can be found in Table 1.2.

- **Chapter 2: Implementations in Machine Ethics: A Survey** — published as paper in ACM Computing Surveys [491]. This chapter investigates the state of the art of implementing ethical theory into AI and answers Research Question 1.
- **Chapter 3: Second Chance for a First Impression? Trust Development in Intelligent System Interaction** — published as paper in the Proceedings of the 29th Conference on User Modeling, Adaptation and Personalization (UMAP '21) [490]. In this chapter, an experiment is presented that researches the impact of system (in)accuracy over time on trust and reliance of users. It answers Research Question 2.
- **Chapter 4: Taxonomy of Trust-Relevant Failures and Mitigation Strategies** — published as paper in the Proceedings of the 15th ACM/IEEE International Conference on Human-Robot Interaction (HRI '20) [492]. A taxonomy of AI failures that influence trust is introduced in this chapter, as well as possible mitigation strategies for each type of error. It answers Research Question 3.
- **Chapter 5: Capable but Amoral? Comparing AI and Human Team Members in Ethical Decision Making** — accepted for Revise and Resubmit at the ACM CHI Conference on Human Factors in Computing Systems (CHI '22). This chapter reports an experiment on user perception of AI for ethical decision making. It answers Research Question 4.

– **Chapter 6: Human-AI Teaming in the Cockpit: Domain Mapping and Research Agenda** — accepted for Major Revision at the 25th ACM Conference On Computer-Supported Cooperative Work And Social Computing (CSCW '22). In this final paper chapter, the state of the art of human-AI teaming is introduced, as well as possible future avenues of research. It answers Research Question 5.
– **Chapter 7: Conclusions.** The final section of this thesis includes chapters for the conclusion, limitations, and proposed future work that builds on the findings of this thesis.

Table 1.2: The contributions per chapter are classified according to Elsevier's *Contributor Roles Taxonomy*.

Contribution	**Chapter 2**	**Chapter 3**	**Chapter 4**	**Chapter 5**	**Chapter 6**
Conceptualization	×	×	×	×	×
Methodology	×	×	×	×	×
Formal analysis	×	×	*n/a*	×	*n/a*
Investigation	×	×	*n/a*	×	*n/a*
Data curation	×	×	*n/a*	×	*n/a*
Writing (original draft)	×	×	×	×	×
Writing (review/editing)	×	×	×	×	×
Visualization	×	×	*n/a*	×	*n/a*
Project administration	×	×	×	×	×

Part II

Contributions to Thesis

Implementing Ethics in AI

This chapter is based on:

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. ***Implementations in Machine Ethics: A Survey****. ACM Computing Surveys, 6, Article 132 (February 2021). DOI: https://doi.org/10.1145/3419633*

Implementations in Machine Ethics: A Survey

Suzanne Tolmeijer[1], Markus Kneer[2], Cristina Sarasua[1], Markus Christen[2], and Abraham Bernstein[1]

[1] Department of Informatics, University of Zurich, Switzerland
[2] Digital Society Initiative, University of Zurich, Switzerland

Abstract. Increasingly complex and autonomous systems require machine ethics to maximize the benefits and minimize the risks to society arising from the new technology. It is challenging to decide which type of ethical theory to employ and how to implement it effectively. This survey provides a threefold contribution. Firstly, it introduces a trimorphic taxonomy to analyze machine ethics implementations with respect to their object (ethical theories), as well as their nontechnical and technical aspects. Secondly, an exhaustive selection and description of relevant works is presented. Thirdly, applying the new taxonomy to the selected works, dominant research patterns and lessons for the field are identified, and future directions for research are suggested.

2.1 Introduction

Autonomous machines are increasingly taking over human tasks. Initially, simple and limited assignments such as assembly line labor were taken over by machines. Nowadays, more complex tasks are transferred to software and robots. Even parts of jobs that were previously deemed purely human occupations, such as being a driver, credit line assessor, medical doctor, or soldier are progressively carried out by machines (e.g., [101, 120]). As many believe, ceding control over important decisions to machines requires that they act in morally appropriate ways. Or, as Picard puts it, "the greater the freedom of a machine, the more it will need moral standards" [387, p. 134].

For this reason, there has been a growing interest in *Machine Ethics*, defined as the discipline "concerned with the consequences of machine behavior towards human users and other machines"[11, p. 1].[3] Research in this field is a combination of computer science and moral philosophy. As a result, publications range from theoretical essays on what a machine can or should do (e.g. [55, 128, 183, 463]), to prototypes implementing ethics in a system (e.g., [8, 524]). In this field, the emphasis lies on how to design and build a machine such that it could act ethically in an autonomous fashion. [4]

The need that complex machines should interact with humans in an ethical way is undisputed; but for understanding which design requirements follow from this necessity requires a systematic approach that is usually based on a taxonomy. There have been several

[3] While there are other terms for the field, such as "Artificial Morality" and "Computational Ethics", the term "Machine Ethic" will be used throughout this survey to indicate the field.

[4] In the following, the expression "implementations in machine ethics" concerns all relevant aspects to successfully create real world machines that can act ethically - namely the object of implementation (the ethical theory), as well as nontechnical and technical implementation aspects when integrating those theories into machines. By "machine" we denote both software and embodied information systems (such as robots).

attempts to classify current approaches of machine ethics. A first high-level classification was proposed by Allen et al. [7] in 2005, distinguishing between top-down theory-driven approaches, bottom-up learning approaches, and hybrids of the two. Subsequent work tried to further determine types of procedures [67, 540, 543], but these works were either mixing different dimensions (e.g., mixing technical approach and ethical theory in one category) [540] or offering an orthogonal dimension that did not fit the existing taxonomy (e.g., whether normative premises can differ between ethical machines) [67]. Also, because these works did not provide an extensive and systematic overview of the application of their taxonomy, verification of the taxonomy with papers from the field was missing. A recent survey from Yu et al. [543] on ethics in AI has some overlap with this work, but 1) does not systematically apply the ethical theory classification to selected papers and 2) takes a broader perspective to include consequences of and interaction with ethical AI, while this paper focuses specifically on machine ethics implementations. Hence, compared to previous works, this survey covers more related work, provides a more extensive classification, and describes the relationship between different ethics approaches and different technology solutions in more depth than previous work [7, 67, 540, 543]. Furthermore, gaps are identified regarding nontechnical aspects when implementing ethics in existing systems.

This paper is created as a collaboration between ethicists and computer scientists. In the context of implementing machine ethics, it can be a pitfall for philosophers to use a purely theoretical approach without consulting computer scientists, as this can result in theories that are too abstract to be implemented. Conversely, computer scientists may implement a faulty interpretation of an ethical theory if they do not consult a philosopher. In such an interdisciplinary field, it is crucial to have a balanced cooperation between the different fields involved.

The contributions of this article are as follows:

- Based on previous work [7], a trimorphic taxonomy is defined to analyze the field based on three different dimensions: types of ethical theory (Section 2.4), nontechnical aspects when implementing those theories (Section 2.5), and technological details (Section 2.6).
- The reviewed publications are classified and research patterns and challenges are identified (Section 2.7).
- An exhaustive selection and description of relevant contributions related to machine ethics implementations is presented (Appendix A).
- A number of general lessons for the field are discussed and further important research directions for machine ethics are outlined (Section 2.8).

As such, this survey aims to provide a guide, not only to researchers but also to those interested in the state of the art in machine ethics, as well as seed a discussion on what is preferred and accepted in society, and how machine ethics should be implemented.

The rest of this paper is structured as follows. Section 2.2 introduces the field of machine ethics, its importance, and justification of used terminology throughout the paper.

Section 2.3 lists the methodology used to create this survey, including the search methodology, the process of creating the classification dimensions, and the actual classification process. Section 2.4, 2.5 and 2.6 introduce the three classification dimensions presented in this survey. Section 2.7 discusses the results of the classification of the selected papers. Finally, Section 2.8 outlines which future avenues of research may be interesting to pursue based on the analysis, as well as the limitations of this survey.

2.2 Introduction to Machine Ethics

Before going into more depth on the implementation of ethics, it is important to establish what is considered machines ethics, why it matters, and present the relevant terminology for the field.

2.2.1 Relevance

Software and hardware (combined under the term "machine" throughout this survey) are increasingly assisting humans in various domains. They are also tasked with many types of decisions and activities previously performed by humans. Hence, there will be a tighter interaction between humans and machines, leading to the risk of less meaningful human control and an increased number of decision made by machines. As such, ethics needs to be a factor in decision making to consider fundamental problems such as the attribution of responsibility (e.g., [463]) or what counts as morally right or wrong in the first place (e.g., [507]). Additionally, ethics is needed to reduce the chance of negative results for humans and/or to mitigate the negative effects machines can cause.

Authors in the field give different reasons for studying (implementations in) machine ethics. Fears of the negative consequences of AI motivate the first category of reasons: creating machines that do not have a negative societal impact [22, 307]. With further autonomy and complexity of machines, ethics need to be implemented in a more elaborate way [59, 110, 140, 240, 348, 380]. Society needs to be able to rely on machines to act ethically when they gain autonomy [11, 139]. A second category of reasons for studying machine ethics focuses on the ethics part: by implementing ethics, ethical theory will be better understood [59, 185, 348, 380].Robots might even outperform humans in terms of ethical behavior at some point [9, 18].

Some authors contend that in cases with no consensus on the most ethical way to act, the machine should not be allowed to act autonomously [10, 463]. However, not acting does not imply the moral conundrum is avoided. In fact, the decision *not* to act also has a moral dimension [177, 251, 530] —think, for example, of the difference between active and passive euthanasia [404]. Additionally, by not allowing the machine to act, all the possible advantages of these machines are foregone. Take, for example, autonomous cars: a large number of traffic accidents could be avoided by allowing autonomous cars on the road. Moreover, simply not allowing certain machines would not stimulate the conversation on

how to solve the lack of consensus; a conversation that can lead to new, more practical ethical insights and helpful machines.

2.2.2 Terminology

An often-used term in the field of machine ethics is "Artificial Moral Agent" or AMA, to refer to a machine with ethics as part of its programming. However, to see if this term is appropriate to use, it is important to identify what moral agents mean in the context of machine ethics and how ethical machines should be regarded. In an often-cited paper, Moor [348] defines four different levels of moral agents:

Ethical-impact agents are types of agents that have an (indirect) ethical impact. An example would be a simple assembly line robot that replaces a human in a task. The robot itself does not do anything (un)ethical by acting. However, by existing and doing its task, it has an ethical impact on its environment; in this case, the human that performed the task is replaced and has to find another job.

Implicit ethical agents do not have any ethics explicitly added in their software. They are considered implicitly ethical because their design involves safety or critical reliability concerns. For example, autopilots in airplanes should let passengers arrive safely and on time.

Explicit ethical agents draw on ethical knowledge or reasoning that they use in their decision process. They are explicitly ethical, since normative premises can be found directly in their programming or reasoning process.

Fully ethical agents can make explicit judgments and are able to justify these judgments. Currently, humans are the only agents considered to be full ethical agents, partially because they have consciousness, free will, and intentionality.

While these definitions can help with a first indication of the types of ethical machines, they do not allow for distinctions from a technical perspective and are also unclear from a philosophical perspective: Moor [348] does not actually define what a moral agent is. For example, it can be debated whether an autopilot is an agent. Therefore, a clearer definition is needed of what an agent is. Himma [230] investigates the concepts of agency and moral agency, drawing from philosophical sources such as the Stanford Encyclopedia of Philosophy and Routledge Encyclopedia of Philosophy. He proposes the following definitions:

Agent : "X is an agent if and only if X can instantiate intentional mental states capable of performing actions." [230, p. 21]

Moral agency : "For all X, X is a moral agent if and only if X is (1) an agent having the capacities for (2) making free choices, (3) deliberating about what one ought to do, and (4) understanding and applying moral rules correctly in the paradigm cases." [230, p. 24]

With regards to artificial agents, Himma postulates that the existence of natural agents can be explained by biological analysis, while artificial agents are created by "intentional agents out of pre-existing materials" [230, p. 24]. He emphasizes that natural and artificial agents are not mutually exclusive (e.g. a clone of a living being). He further claims that moral agents need to have conscious and intentionality, something that state-of-the-art systems do not seem to instantiate. It is worth noting that Himma attempts to provide a general definition of moral agency, while for example Floridi and Sanders [173] propose to change the current description of a moral agent. For example, they proposed description includes the separation the technical concepts of moral responsibility and moral accountability, a distinction that was not evident thus far: "An agent is morally accountable for x if the agent is the source of x and x is morally qualifiable [...] To be also morally responsible for x, the agent needs to show the right intentional states". Wallach and Allen [513] rate AMAs along two dimensions: how sensitive systems are to moral considerations and how autonomous they are. Sullins [473] has a partially overlapping concept of requirements for robotic moral agency with Himma's, that intersects with Wallach and Allen's relevant concepts: autonomy (i.e. "the capacity for self-government") [63]), intentionality (i.e. "the directedness or 'aboutness' of many, if not all, conscious states"[63]) and responsibility (i.e. "those things for which people are accountable"[63]).

These are just some notions of how concepts such as agency, autonomy, intentionality, accountability and responsibility are important to the field of machine ethics. However, it is challenging to summarize and define these concepts concisely while doing justice to the work in philosophy, and computer science that has been done so far, including the discussions and controversy around different relevant concepts (as the different concepts of moral agency display). The goal of this survey is not to give an introduction to moral philosophy, so this section merely gives a glimpse of the depths of the topic. Rather, the goal is to summarize and analyze the current state of the field of machine ethics. To avoid any assumption on concepts, the popular term Autonomous Moral Agent is not used in this survey: as shown above, the term "agent" can be debated in this context and the term "autonomous" has various meanings in the different surveyed systems. Instead, a machine that has some form of ethical theory implemented—implicitly or explicitly—in it is referred to as an "ethical machine" throughout this paper. Accordingly, we refrained from adding an analysis regarding degree of agency and autonomy of machines into our taxonomy, as those points are rarely discussed by the authors themselves and because they would have added a layer of complexity that would have made our taxonomy confusing.

2.3 Survey Methodology

This section describes the search strategy, paper selection criteria, and review process used for this survey.

2.3.1 Search Strategy

A literature review was conducted to create an overview of the different implementations of and approaches to machine ethics. The search of relevant papers was conducted in two phases: automated search and manual search.

Automated Search The first phase used a search entry that reflected different terms related to machine ethics combined with the word 'implementation':

> implementation AND ("machine ethics" OR "artificial morality" OR "machine morality" OR "computational ethics" OR "roboethics" OR "robot ethics" OR "artificial moral agents")

These terms were cumulated during the search process (e.g. [515, p. 455]); each added term resulted in a new search until no new terms emerged.[5] No time period of publication was specified, to include as many items as possible.

The following library databases were consulted (with the number of results in parenthesis): Web of Science (18), Scopus (237), ACM Digital Library (16), Wiley Online Library (23), ScienceDirect (48), AAAI Publications (4), Springer Link (247), and IEEE Xplore (113). Of these initial results, 37 items were selected based on the selection criteria listed in Section 2.3.2.

Manual Search The second phase included checking the related work and other work by the same first authors of phase one. Twenty-nine promising results from phase one did not meet all criteria, but were included in the second search phase to see if related publications did meet all criteria. This process was repeated for each newly found paper, until no more papers could be added that fit the selection criteria (see 2.3.2). This resulted in a total of 49 papers, describing 48 ethical machines.

2.3.2 Selection Criteria

After the selection process, two more coauthors judged which papers should be in- or excluded to verify the selection. Papers were included only if they adhered to all of the following inclusion criteria. The paper

- implements a system OR describes a system in sufficient (high-level) detail for implementation OR implements/describes a language to implement ethical cases,
- describes a system that is explicitly sensitive to ethical variables (as described by [348]), no matter whether it achieves this sensitivity through top-down rule-based approaches or bottom-up data-driven approaches (as described by [7]),

[5] The term "Friendly AI", coined by Yampolsky [544], is excluded since it describes theoretical approaches to machine ethics.

- is published as a conference paper, workshop paper, journal article, book chapter, or technical report,
- and has ethical behavior as the main focus.

The following exclusion criteria were used. The paper

- describes machine ethics in a purely theoretical fashion,
- describes a model of (human) moral decision making without an implementable model description,
- lists results of human judgment on ethical decisions without using the data in an implementation,
- is published as a complete book, presentation slides, editorial, thesis, or has not been published,
- describes a particular system in less detail than other available publications,
- focuses on unethical behavior to explore ethics (e.g., a lying program),
- mentions ethical considerations while implementing a machine, but does not focus on the ethical component and does not explain it in enough detail to be the main focus,
- simulates artificial agents to see how ethics emerge (e.g. by using an evolutionary algorithm without any validation),
- and describes a general proposal of an ethical machine without mentioning implementation related details.

Given the focus on algorithms implementing moral decision making and the limitations of space, we will not go into further detail as regards recent interesting work on AI and moral psychology (cf. e.g. [34, 66, 301, 440]).

2.3.3 Taxonomy Creation and Review Process

In order to be able to identify strengths and weaknesses of the state of the art, we created different taxonomies and classified the selected papers accordingly. It was clear that both a dimension referring to the implementation object (the ethical theory; cf. Table 2.1) and a dimension regarding the technical aspects of implementing those theories (cf. Table 2.4) were necessary. All authors agreed that there were some aspects of implementing those theories that did not concern purely technical issues but were still important for the classification. Hence, we defined and applied a third taxonomy dimension (cf. Table 2.3) related to non-technical implementation choices. The first version of these three taxonomies was created using knowledge obtained during the paper selection.

Before the classification process started, one third of the papers were randomly selected to review the applicability of the taxonomy proposed and adjust the assessment scheme where required. Any parts of the taxonomies that was unclear and led to inconsistent classifications was adapted and clarified.

Each selected paper was categorized according to the features of the three different taxonomy dimensions (discussed in Sections 2.4–2.6). Since the ethical classification is perhaps the most disputable, it was determined by three distinct assessors: two philosophers and a computer scientist. Two computer scientists evaluated the implementation and technical details of all proposed systems.

To provide a classification for all selected papers, multiple classification rounds took place for each dimension. In between classification rounds, disagreements across reviewers were discussed until a consensus was reached. In the case of the ethical dimensions, four papers could not be agreed upon after multiple classification rounds. As such, these papers were labeled as 'Ambiguous'.

Additionally, we shared a pre-print of the article with the authors of the classified systems in order to verify that they agreed with the classifications we provided. From 45 targeted authors, we received 18 responses. From these 18, 6 authors agreed with our classification and 12 proposed (mostly minor) changes or additional citations. In total, we changed the classification of 6 features for 4 papers.

In the following, we now outline the three dimensions of our taxonomy.

2.4 Object of Implementation: Ethical theories

This section introduces the first of three taxonomy dimensions introduced in this paper: a taxonomy of types of ethical theories, which is the basis for the categorization of ethical frameworks used by machines (in Section 2.7). Note that this section is not a general introduction to (meta-)ethics, which can for example be found in [62, 113, 343].

2.4.1 Overview of Ethical Theory Types

It is commonplace to differentiate between three distinct overarching approaches to ethics: **consequentialism**, **deontological ethics**, and **virtue ethics**. Consequentialists define an action as morally good if it maximizes well-being or utility. Deontologists define an action as morally good if it is in line with certain applicable moral rules or duties. Virtue ethicists define an action as morally good if, in acting in a particular way, the agent manifests moral virtues. Consider an example: an elderly gentleman is harassed by a group of cocky teenagers on the subway and a resolute woman comes to his aid. The consequentialist will explain her action as good since the woman maximized the overall well-being of all parties involved—the elderly gentleman is spared pain and humiliation which outweighs the teenagers' amusement. The deontologist will consider her action commendable as it is in accordance with the rule (or duty) to help those in distress. The virtue ethicist, instead, will deem her action morally appropriate since it instantiates the virtues of benevolence and courage.

Consequentialist theories can be divided into two main schools: according to *act utilitarianism*, the principle of utility (maximize overall well-being) must be applied to each

individual act. *Rule utilitarians*, by contrast, advocate the adoption of those and only those moral rules that will maximize well-being. Cases can thus arise where an individual action does not itself maximize well-being, yet is consistent with an overarching well-being maximizing rule. While act utilitarians would consider this action morally bad, rule utilitarians would consider it good.

Deontological ethics can be divided into agent-centered and patient-centered approaches. *Agent-centred theories* focus on agent-relative duties, such as, for instance, the kinds of duties someone has towards their parents (rather than parents in general). Theories of this sort contrast with *patient-centered theories* that focus on the rights of patients (or potential victims), such as the right, postulated by Kant, not to be used as a means to an end by someone else [264].

Finally, there are some approaches that question the universal applicability of general ethical principles to all situations, as put forward by deontological ethics, virtue ethics or consequentialism. For such a **particularist view**, moral rules or maxims are simply vague rules of thumb, which cannot do justice to the complexity of the myriad of real-life situations in which moral agents might find themselves. Hence, they have to be evaluated on a case-by-case basis.

We would like to highlight that a moral theory is a set of substantial moral principles that determine what, according to the theory, is morally right and wrong. Moral theories can take different structures —they might state their concrete demands in terms of hard rules (deontological ethics); virtues that should guide actions, with reference to an overall principle of utility maximization, or else reject the proposal that there is a one-size-fits-all solution (itself a structural trait, this would be particularism). In this work, we are interested in these structures, which we label "ethical theory types".

2.4.2 Categorizing Ethical Machines by Ethical Theory Type

Based on the distinct types of ethical theories introduced above, this sub-section develops a simple typology of ethical machines, summarized in Table 2.1.

An evaluation of existing approaches to moral decision making in machines can make use of this typology in the following way. Deontological ethics is rule-based. What matters is that the agent acts in accordance with established moral rules and/or does not violate the rights of others (whose protection is codified by specified rules). Accidents occur, and a well-disposed agent might nonetheless bring about a harmful outcome. On off-the-shelf deontological views, bad outcomes (if non-negligently, or at least unintentionally, brought about) play no role in moral evaluation, whereas the agent's mental states (their intentions, , and beliefs) are important. If John, intending to deceive Sally about the shortest way to work, tells the truth (perhaps because he himself is poorly informed), a Kantian will consider his action morally wrong, despite its positive consequence.[6] In the context of machine ethics, the focus is solely on agent relative duties. Hence, no distinction

[6] Note that if two actions differ only with respect to outcome, consequences can play a role.

is made between agent-centered and patient-centered theories of deontological ethics in the taxonomy summarized in Table 2.1.

Consequentialists, by contrast, largely disregard the agent's mental states and focus principally on outcomes: what matters is the maximization of overall well-being. Note that, procedurally, a rule-utilitarian system can appear very similar to a deontological one. The agent must act in keeping with a set of rules (potentially the very same as in a Kantian system) which, in the long run, maximizes well-being. However, the two types of systems can still be distinguished in terms of the ultimate source of normativity (well-being vs. good will) and will—standardly—differ in terms of the importance accorded to the agent's mental states. Thus far, nearly all consequentialist machine ethics implementations utilize act utilitarianism. For this reason, the distinction between act and rule utilitarianism is not relevant enough to be included in this survey.

Table 2.1: Ethical theory types
taxonomy dimension

Ethics Type
Deontological ethics
Consequentialism
Virtue ethics
Particularism
Hybrid
- *Hierarchically specific*
- *Hierarchically nonspecific*
Configurable ethics
Ambiguous

Virtue ethics differs from the aforementioned systems in so far as it does not focus principally on (the consequences or rule-consistency of) actions but on agents, and more particularly on whether they exhibit good moral character or virtuous dispositions. A good action is one that is consistent with the kinds of moral dispositions a virtuous person would have.

In contrast to the other three major approaches, on the particularist view, there is no unique source of normative value, nor is there a single, universally applicable procedure for moral assessment. Rules or precedents can guide our evaluative practices. However, they are deemed too crude to do justice to many individual situations. Thus, according to particularism, whether a certain feature is morally relevant or not in a new situation —and if so, what exact role it is playing there— will be sensitive to other features of the situation.

Table 2.2 gives a schematic overview of key characteristics of the different types of ethical systems that might be implemented in an ethical machine. Note that it does not take some of the more fine-grained aspects differentiating the theories (e.g., the beforementioned complications regarding act and rule utilitarianism) into account.

As an alternative to implementing a single determinate type of ethics, systems can also combine two or more types, resulting in a **hybrid** ethical machine. This approach seems enticing when one theory alleviates problems another one might have in certain situations, but it can also generate conflicts across types of ethical approaches. Hence, some proposals enforce a *specified hierarchy*, which means that one theory is dominant over the other(s) in the system. For example, a primarily deontological system might use rules, but turn to the utilitarian approach of maximizing utility when the rules are in

Table 2.2: High-level overview to ethics categories in the context of ethical machine implementation

	Input	Decision criteria	Mechanism	Challenges (examples)
Deontological ethics	Action (mental states and consequences)	Rules/duties	Fittingness with rule	• Conflicting rules • Imprecise rules
Consequentialism	Action (consequences)	Comparative well-being	Maximization of utility	• Aggregation problems • Determining utility
Virtue ethics	Properties of agent	Virtues	Instantiation of virtue(s)	• Conflicting virtues • Concretion of virtues
Particularism	Situation (context, features, intentions, consequences)	Rules of thumb, precedent, all situations are unique	Fittingness with rules/precedent	• No unique and universal logic • Each situation needs unique assessment

conflict. In other cases, the hierarchy is *non-specific* and different theories are present without a specified dominant theory.

Some authors do not settle on a particular type of ethical theory. Instead, they provide a **configurable** technical framework or language and exhibit how different types of ethical theories can be implemented. The choice of which theory type should be selected is essentially left to the person implementing the system in an actual use case.

Finally, some contributions were classified as **ambiguous** from a meta-ethical perspective. For these, not enough details were given by the authors to classify a paper, or the theories used to implement were not ethical theories but retrieved from domains other than moral philosophy.

2.4.3 Ethical Theory Types in Practice

There are certain challenges inherent in the different types of ethics when they need to be applied in practice. Since these obstacles need to be taken into account to select an ethical theory type for an ethical machine, this subsection provides a (non-exhaustive) list of complications.

Challenges of deontological ethics in practice: At a first glance, the rule-based nature of deontological ethics seems to lend itself well for implementation. However, at different stages of implementation, challenges arise. The first issue is *which rules should be implemented*. Rules are expected to be strictly followed, implying that for every exception, the rule must be amended, resulting in an extremely long rule. Determining the right level of detail is important for the success of an application: when the rules are not practical and at the right level of detail, they will not be interpretable for the machine [18]. Second, there might be *conflicts between rules* [67]—in general or in specific situations. Whilst ordering or weighing the rules might address this issue from an implementational perspective, determining an order of importance can be difficult. Also, this assumes that all relevant rules are determined before they are used.

Challenges of consequentialist ethics in practice: There are three main categories of difficulties for consequentialist ethics. First, it is hard to identify consequences and determine the right level of detail and aggregation in terms of time and size. Some outcomes might have resulted regardless of the action theorized to have caused it. In real-life situations, all possible consequences are not always that clear beforehand given the lack of epistemic transparency and causal interdependence.

A second issue is concerned with quantifying consequences. As consequentialism is about maximizing utility, the problem is how to define *utility*. In simple scenarios like the Trolley problem, utility is often defined as how many people survive or die. In the real world, more complex concepts, such as happiness and well-being, are preferred to define utility. There are measures available (e.g., QALY [201]), but using a different measure can give a different outcome. Even more so, even if each consequence is assigned a utility, it might still be inappropriate to simply aggregate them (e.g., see [277]).

Finally, there might be a significant computational cost when computing utility [522] requiring heuristics or approximations to derive a correct answer in time. This, in turn, requires a verification of whether these results are still correct.

Challenges of virtue ethics in practice: Virtues are positive character traits, character traits that should be manifested in morally good actions. Defining what "character" a machine has is troubling, if a machine can be claimed to have a character at all. To judge whether a machine—or a human for that matter—is virtuous is not possible by merely observing one action or a series of actions that seem to imply that virtue; the reasons behind them need to be clear [466]. Perhaps the best way to create a virtuous machine is to let a machine mimic the behavior of a virtuous person. But how is a certain virtue measured, and who decides which virtues are more important and how to pick the perfect role model? Coleman [111] even proposes different virtues that are more desirable for machines rather than human virtues, implying merely mimicking a virtuous person is not sufficient.

To circumvent these challenges, machine ethics researchers have not used virtue ethics often, as the alternatives might be more appealing. For example, Haber [203] states that virtue ethics and principle-based ethics are complements and that for each trait there will be a principle that expresses that trait and vice versa. While not everyone agrees with Haber, it is easier and more detailed from a computational perspective to implement rules than generic virtues to adhere to. Arkin [18] also concludes that principle-based and act-centric models allow for stricter ethical implementations, which is desirable in machine ethics.

Challenges of particularism in practice: In particularism, the system needs to take the entire context into account. This implies that it needs to either be trained for all possible cases, which is not possible, or be able to extrapolate without using generalizations, which is highly challenging. For each feature of the context, the system would have to recog-

nize whether it is morally relevant in the given case and how it will influence the result. Case-based methods or instance-based classifications come closest to allowing an implementation of particularism. More recently, some contributions are trying to approximate particularist ethics using neural networks (e.g., [199, 216]).

Challenges of hybrid approaches in practice: Each type of ethical theory raises its own set of complications, but combining them introduces additional issues. First, when different types of ethical theories are used in a non-hierarchical way, the interaction between them can be problematic: how should the results from different ethical approaches be combined to guarantee morally appropriate outcomes? What happens when the results of different implemented ethical theory types stand in conflict, and how should such conflicts be resolved?

Second, when a hierarchical approach is employed, it is not evident when the system should employ one theory rather than another. One standard approach resorts to the secondary set of ethical principles when the first does not deliver a verdict. While this alleviates some of the challenges of hybrid systems, it is still possible that the second ethical theory proposes something that conflicts with the first ethical theory type.

The next section introduces the second dimension of ethical machines: the non-technical aspects of implementing ethics into a machine.

2.5 Non-Technical Implementation Aspects

The second taxonomy dimension that was created for this survey considers the non-technical aspects of implementing the aforementioned ethical theories into machines. An important part of creating an ethical system is to decide *how* to implement ethics. That entails defining whether an implementation can follow different approaches, how to evaluate the system, and whether or not domain specifications need to be taken into account. Important features concerning the implementation dimension are summarized in Table 2.3. Furthermore, this section highlights the implementation challenges that the various ethical theories entail.

2.5.1 Approaches

Different typologies have been proposed to determine how ethics types are implemented. The most influential and widely referenced scheme, also applied in this survey, stems from Allen, Smit and Wallach [7]. They distinguish three types of implementation approaches, namely top-down, bottom-up, and hybrid.

Top-down approaches : Top-down approaches assume that humans have gathered sufficient knowledge on a specific topic; it is a matter of translating this knowledge into an implementation. The ethical theory types described in Section 2.4 are examples of normative human knowledge that can be translated into usable mechanisms for machines.

Table 2.3: Non-technical taxonomy dimension

Feature	**Type**	**Subtype**
Approach	Top-down Bottom-up Hybrid	
Diversity consideration	Yes No	
Contribution type	Model representation Model selection Judgment provision Action selection/execution	
Evaluation	Test	Non-expert Expert Laws
	Prove	Model checker Logical proof
	Informal	Example scenarios Face validity
	None	
Domain specific	Yes (*domain specified*) No	

The system acts in line with predetermined guidelines and its behavior is therefore predictable. In AI, strategies using a top-down approach mostly make use of logical or case-based reasoning. Given general domain knowledge, the system can reason about the situation that is given as input. Usually, human knowledge is not specified in a very structured or detailed way for concrete cases, so knowledge needs to be interpreted before it can be used. This process presents the risk of losing or misrepresenting information. The positive aspect of this approach is that existing knowledge is applied and no new knowledge needs to be generated.

Bottom-up approaches : A different method to implementing ethics is to assume the machine can learn how to act if it receives as input enough correctly labeled data to learn from. This approach, not just in machine ethics but in general, has gained popularity after the surge of machine learning in AI and the recent success of neural networks. Technologies such as artificial neural networks, reinforcement learning, and evolutionary computing fall under this trend. Increased computing power and amounts of data allow learning systems to become more successful. However, data has to be labeled consistently and the right data properties need to be described in a machine-processable way to obtain an accurate training of machines. There is a risk that the machine learns the wrong rules or cannot reliably extrapolate to cases that were not

reflected in its training data. However, for certain tasks, such as feature selection or classification, machine learning can be very successful.

Hybrid approaches : As the term suggests, hybrid approaches combine top-down and bottom-up approaches. As Allen et al. phrase it: "Both top-down and bottom-up approaches embody different aspects of what we commonly consider a sophisticated moral sensibility." [7, p 153] They indicate that a hybrid approach is considered necessary, if a single approach does not cover all requirements of machine ethics. The challenge consists in appropriately combining features of top-down and bottom-up approaches.

Bonnemains et al. [67] suggest adding a fourth category, called "Personal values/ethics system". Essentially, it acknowledges that two different agents may rely on different ethical systems or may rely on different precedence in case of conflicts in a hybrid system. In this survey, this is regarded as **diversity consideration**: the authors of a machine ethics paper consider the possibility that not all ethical machines adhere to the same ethical theory type, and their contribution includes the choice of diverse types of ethics to be implemented. As Bonnemains et al. recognize, this category is somewhat orthogonal to the previous three, as all of those can be seen to implement distinct normative principles. For example, a machine ethics implementation with diversity consideration could allow for multiple ethical theory types to be implemented (i.e., a top-down approach) or allow for different machines to learn different types of ethics (i.e., a bottom-up approach). It is considered part of the implementation dimension rather than the ethics dimension since diversity considerations can also exist within the same ethical theory, for example, by allowing deontological machines to have different rules to adhere to while still all being deontological in nature. This survey regards structures of normative frameworks and their implementation rather than substantial normative principles (c.f. Table 2.3).

2.5.2 Type of Contribution

Ethical systems can be intended to enact different aspects of ethical behavior. This section discusses the different types of contributions published to implement ethical machines.

Model representation : This contribution type focuses on representing current ethical knowledge. The goal is to determine how to appropriately represent a theory, dilemma, or expert-generated guidelines whilst staying true to the original theory.

Model selection : Given a set of alternative options to implement an ethical machine, some systems limit their action to selecting the most fitting elements to be included in the system.

Judgment provision : These contributions focus on judging an action given a scenario and a set of possible actions. Example outputs are binary (*acceptable*/*non-acceptable*) or responses on a scale (e.g., *very ethical* to *very unethical*).

Action selection/execution : Here the proposed system chooses which action is best given multiple possible actions for a scenario. Some systems then assign the action

to a human, while others carry out the selected action themselves. Part of the action selection task can also be action restriction, when some possible actions are not morally acceptable (enough).

2.5.3 Evaluation

Most artifacts—simple or complex, concrete or abstract—can be evaluated in virtue of their capacity to fulfill their constitutive function or purpose. A good knife cuts well, a good thermostat reliably activates the heating if the temperature drops below a predetermined threshold, and a good translation system adequately and idiomatically converts grammatical sentences from one language into another. Whereas there are objective and measurable criteria for the evaluation of thermostats, things are more cumbersome when it comes to moral machines. This is not because their purpose does not standardly consist in simply "acting morally", but in executing certain tasks (taking care of the elderly, counselling suicidal people, evaluating risk of recidivism etc.) in a moral fashion. Much rather, the complication arises from the question of what exactly is to count as executing the task at hand in morally appropriate ways, or against what exactly the behavior of the system should be evaluated.

There are objective facts as to whether an image represents a certain type of animal or not. These facts constrain whether the image is correctly classified as representing an animal. The existence of objective, universal moral values, by contrast, is controversial (cf. e.g., [217, 319, 401, 503]). Furthermore, and as objectivists readily acknowledge, delineating what is morally permissible poses an epistemic challenge of a different order than identifying, say, a giraffe in an image, or determining the weight of an object. The ontological and epistemic complications that arise in the moral domain thus make it difficult to settle on standards against which the performance of a moral machine could be evaluated. More fundamentally, it is not even evident what kinds of considerations should guide the process of choosing such standards.

While complications as to the evaluation of a moral machine are worrying, their practical significance should not be exaggerated. Although there is disagreement as regards complex cases, in ordinary life situations in which one is confronted with extremely difficult ethical decisions or run away trolleys are exceedingly rare. In many domains, moral dilemmas are unlikely to arise or be of much import, and there is widespread convergence (not only among the folk, but experts, too) on what constitutes adequate moral behavior. Overall, then, the challenge of evaluation might raise metaphysical and epistemic complications of limited pragmatic importance, at least when care is exercised to limit the decision capacity of moral machines to mundane contexts that steer clear of complex ethical paradoxes.

2.5.3.1 Test When a system is tested, the system outcome needs to be compared against a ground truth. These may have the following origins:

Non-experts : One possibility consists in making folk morality the benchmark. Problematically, there is substantial evidence of moral parochialism across cultures (e.g., [167, 318, 425]), and it is not difficult to find topics on which a single nation is roughly divided —just think of abortion, euthanasia, or same-sex relations in the US [423]. Furthermore, the existence of widespread convergence in moral opinion does not necessarily make such opinions true or acceptable (consider that until a century and a half ago, there was broad agreement in considerable parts of the world that slavery is morally acceptable).

Experts : To escape the tyranny of a potentially mistaken or self-serving majority, one might adopt the standard of experts in normative ethics. Problematically, however, experts themselves are sometimes deeply divided on fundamental issues of moral import as well as meta-ethical intuitions [69] and their very expertise can be called into question [445, 446].

Laws : One might side-step the complications raised by retreating to a second-best solution: the law. This strategy, however, is not without drawbacks either, as the law is simply silent on most questions of day-to-day morality. It is, for instance, not illegal to lie in most contexts, yet it would be regarded as outrageous to be perpetually deceived by "moral" machines. Still, it might be suitable to draw on the law to provide restrictions where they exist, for example, as concerns the "Laws of War" or "Laws of Armed Conflicts" for the lethal weapons domain [18], or specific domain rules such as the Code of Ethics for Engineers [336]. As Arkin [18] suggests, scoping the problem using domain-specific requirements can make it more easily implementable and testable.

2.5.3.2 Prove Another approach, typically based on some type of logic, consists of proving that the system behaves correctly according to some known specifications. This approach can be divided into the following types:

Model checker : Given an ethical machine, a model checker exhaustively and automatically ascertains that it adheres to a given set of specifications.

Logical proof : This approach provides a logical proof that given certain premises, the system does what it should do. Proofs of this sort can be effected manually, or by using a theorem prover which employs automated logical and mathematical reasoning.

Note that this approach assumes that a correct specification exists a priori and is widely accepted. Within the logic community, model checking and theorem proving are often considered an implementation issue rather than a type of evaluation (e.g., see [208]). In some cases, authors do not even explicitly mention that they employ a model checker, because it is inherent in their approach to logic programming. However, given the multidisciplinary nature of the field of machine ethics, it is vital to explicitly state which approach has been used. Furthermore, while logical/internal validity and consistency may be inherent in the system, a form of evaluation is necessary to ensure the system acts as expected in different cases and exhibits external validity.

2.5.3.3 Informal evaluation Some authors refrain from formally evaluating their implementation. Instead, they only describe their work and, in some cases, show a few example scenarios or exhibit application domains. Whilst these approaches may have limited validity, they may be warranted given the evaluation complications outlined above or when the authors principally engage in theory building [152].

Example scenarios/case studies : To showcase that the system works as intended, one or multiple scenarios are presented to demonstrate the system's performance. This procedure gives a first indication of the functionalities of the machine or may help in theorizing about certain properties of a system, but it does not cover all possible situations or give a complete performance indication.

Face validity : Often described as "reasonable results", authors using this approach state that the results of a few example tasks are as expected. It is often unclear what this means and to what extent these results are desirable.

2.5.3.4 None When no evaluation could be discerned, papers were categorized as having none of the evaluation types present.

2.5.4 Domain Specificity

What is deemed an appropriate action can depend on the domain in which the moral agent is operating, such as the principles in the domain of biomedical ethics as proposed by Beauchamps and Childress [49] for the medical domain, or the Rules of Engagement and Laws of Armed Conflict for autonomous weapon systems [18]. Hence, some contributions focus on a specific application domain, which limits the scope of an ethical machine implementation, and thus the endeavor is more manageable [18].

2.6 Technical Implementation Aspects

The third and final taxonomy dimension introduced concerns the technical aspects when implementing ethical theories into machines. This includes the type of technology chosen for the implementation, the input the system relies on, the ethical machine's availability (i.e., implementation details are published) and other technical features: whether it relies on specific hardware or feedback from users, provides explanations for its conclusions, has a user interface (UI), and whether the input for the system needs to be preprocessed. Important features pertaining to the technical dimension are surveyed in Table 2.4.

Table 2.4: Technical taxonomy dimension. As explained in Section 2.6, "Inductive logic" is present twice.

Feature	Type	Subtype or classification scheme
Tech type	Logical reasoning	Deductive logic
		Non-monotonic logic
		Abductive logic
		Deontic logic
		Rule-based system
		Event calculus
		Knowledge representation & Ontologies
		Inductive logic
	Probabilistic reasoning	Bayesian approach
		Markov models
		Statistical inference
	Learning	Inductive logic
		Decision tree
		Reinforcement learning
		Neural networks
		Evolutionary computing
	Optimization	
	Case-based reasoning	
Input	Case	Logical representation
		Numerical representation
		(Structured) language representation
	Sensor data	
Implementation availability	Specification details	Y - P - N (Yes - Partially - No)
	Implementation details	Y - P - N
	Code (link) provided	Y - P - N
Other	Hardware (simulation)	Y - P - N
	Feedback	Y - P - N
	Explanation	Y - P - N
	UI(mostly GUI)	Y - P - N
	Automated processing	Y - P - N

2.6.1 Types of Technology

Inspired by Russell and Norvig [421], different types of technologies can be distinguished. While these types of technology are not always clearly delimited, this categorization allows comparing implementations.

2.6.1.1 Logical reasoning There are different types of logic or logic-based techniques used in machine ethics.

Deductive logic : This is the classical type of logic: knowledge is represented as logical statements—propositions and rules—that allow deriving new propositions. Pure deductive systems typically involve no learning or inference involved but only derive what can be known from their set of statements and inputs.

Non-monotonic logic : Non-monotonic logic allows the revision of conclusions when a conflict arises, for example, in light of new information.

Abductive logic : In abductive logic, the conclusions drawn are the most likely propositions given the premises.

Deontic logic : This type of logic stems from philosophy and is specifically designed to express normative propositions. Naturally, this type of logic is inherently suited for the representation and deduction of moral propositions.

Rule-based systems : As the name suggests, rule-based systems are systems that function based on a set of rules. These can be ethical rules the system has to adhere to. Note that many of the different types of logic above are typically implemented as some form of rule-based system.

Event calculus : Event calculus allows reasoning about events. When a machine needs to act ethically, different events can trigger different types of behavior.

Knowledge representation (KR) and ontologies : A KR approach focuses on representing knowledge in a form that a computer system can utilize. In other words, the emphasis lies on improving the quality of the data rather than (just) improving the algorithm.

Inductive logic : When relying on inductive logic, premises are induced or learned from examples, rather than pre-defined by a human.

2.6.1.2 Probabilistic reasoning Recently, probabilistic reasoning has gained more attention. Different types of probabilistic reasoning approaches can be distinguished.

Bayesian approaches Based on Bayes' rule, these approaches rely on prior knowledge to compute the likelihood of an event. In an ethical context, a machine can then act based on this predicted information.

Markov models Markov models focus on sequences of randomly changing events, assuming that a future event only depends on the current (and not the previous) event(s).

Statistical inference By retrieving probability distributions from available data, the system can try to predict the chances of future events happening.

2.6.1.3 Learning The increased computational power, the amounts of data available, and the GPU-driven revival of neural networks have made learning systems more popular. There are different learning approaches to be characterized.

Inductive logic : In inductive logic, a rule-base for reasoning is learned. As such, it is listed under both the "Logic" and "Learning" categories of this taxonomy.

Decision tree : Decision trees are a supervised learning method to solve a classification problem by exploring the decision space as a search tree and computing the expected utility. They are, thus, useful to identify and interpret the features that are most important to classify cases.

Reinforcement learning : A system can learn from its actions when they are reinforced with rewards or punishments received from its environment.

Neural networks : A neural network can be trained on many cases, to be able to classify new cases based on their relevant features.

Evolutionary computing : Evolutionary algorithms are used when, for example, different competing models of an ethical machine exist. Models evolve in an iterative fashion, based on actions inspired from the concept of evolution in the field of biology (e.g., selection, and mutation) [253]

2.6.1.4 Optimization The most common form of optimization relies on a closed-form formula for which some optimal parameters are sought. Different actions get assigned different values based on a predetermined formula, and the best value is chosen (e.g., the highest value).

2.6.1.5 Case-based reasoning In case-based reasoning, a new situation is assessed based on a collection of prior cases. Similar cases are identified and their conclusions are transferred to apply to the current situation.

2.6.2 Input

To be able to respond appropriately, ethical machines need to receive information about the environment (or situation at hand). Input is the information that the system receives, not the transformation the system itself performs on the data afterwards.

Sensor data : In the case of (simulated) hardware, the machine perceives the input through its sensors. The sensor data is interpreted and processed to serve as the input.

Case: Logical representation : Systems using a form of logic often need an input case represented using logic.

Case: Numerical representation : Other systems, for example ones using neural nets, need their input in a numerical form. This can be a vector representation or a set of numbers.

Case: Language representation : Language inputs can be natural language or input translated into structured language.

2.6.3 Implementation availability

As mentioned before, part of the field of machine ethics tends to be of a theoretical nature. This becomes apparent in the level of detail of implementation proposals. While

some authors implement an idea and provide the source code, this is fairly rare in machine ethics. Some authors only give a few implementation details, and others merely specify a high-level description of their idea. Usually, the focus lies on sketching an idea rather than its complete implementation.

Specification details : This level has the fewest implementation details: the author specifies the proposed idea (e.g., textually) without any additional detail.

Implementation details : This next level provides implementation details illustrating how the specification is implemented in the described machine.

Code (link) provided : This final level provides the link to the code of the machine, so the prototype can be used and the experiments can be replicated.

2.6.4 Other Implementation Categories

This section introduces different and independent categories that are of interest for the implementation of an ethical machine.

2.6.4.1 Hardware Robots can have direct physical results rather than "just" digital or indirect physical consequences. Hardware can change the way people interact with a system and how it should be able to function, making it an interesting and important feature to classify.

2.6.4.2 Feedback No matter which ethical approach is used, feedback is a valuable component of an ethical system. For example, the user can be asked whether the provided output was the best given the input or whether the system was clear during its decision process.

2.6.4.3 Explanation Transparency is important when it comes to algorithmic decisions, both from a user perspective [545] and, in some cases (such as the General Data Protection Regulation in the European Union [191]), from a legal perspective. To achieve this goal, an understandable explanation should be provided by the system.

2.6.4.4 User Interface (UI) Systems should be easy to interact with. This is important for all machines, including ethical machines.

2.6.4.5 Automated processing Sometimes, initial prototypes focus on the concept of a system, not the (detailed) implementation, and may require some pre-processing of the input data. Ideally, systems should be able to process input from the environment automatically.

2.7 Analysis

The goal of this section is to classify the surveyed moral machines according to the three taxonomy dimensions introduced in Sections 2.4–2.6 and elicit patterns in the literature based on this classification. Specifically, every publication is categorized according to the object of implementation (ethical theory) as well as the non-technical and technical aspects of implementing those theories (as described in Section 2.3). Summaries of the selected papers can be found in Appendix A.

2.7.1 Ethical Classification

The classification results for the ethical dimension of machine ethics implementations can be found in Table 2.5; the ratio of single vs. hybrid theory papers is visualized in Figure 2.1. Among the papers, several constitute clear-cut cases instantiating one of the four main ethical systems. For example, [10, 359, 455] are clearly deontological; and [2, 106, 140, 502] constitute uncontroversial examples of consequentialist systems. Furthermore, a considerable number of papers invoke elements from multiple systems. Finally, there are papers in which the hierarchy across theory types remains ambiguous. Examples of ambiguous papers are implementations where authors try to mimic the human brain [90], or focus on implementing constraints such as the Pareto principle [307], which does not strictly speaking constitute a moral theory. Note that categorizing a paper as "ambiguous" does not imply a negative assessment of the implementation. It simply means that the proposal cannot be adequately placed within our classification framework.

About 50% of the proposals draw on a single type of ethical theory (see Figure 2.1). As can be seen in Table 2.5, deontological and consequentialist ethics are used most often. It stands out that particularism is barely used and pure virtue ethics is not used at all. This may be explained as follows: first, a generalist approach is much easier to implement than a particularist approach, as it is more straightforward to encode generalist rules than to build systems that may have to handle as of yet unknown, particular cases. Second, virtue ethics can be considered a very high-level theory focusing on characteristics rather than actions or consequences, which is difficult to interpret in an application context.

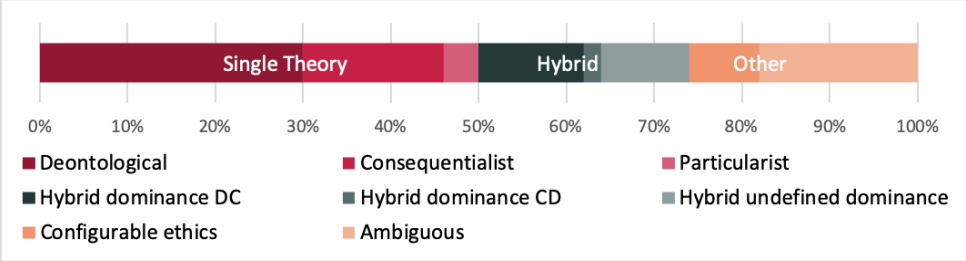


Fig. 2.1: Ethical theory type ratio

Table 2.5: Ethical theory classification. *Hybrid dominance D-C* implies both D and C are implemented, but D is dominant. The reverse is true for *Hybrid dominance C-D*. For the *Hybrid undefined dominance* the theories that are combined are noted in parentheses following the citation.

Ethical theory type	Papers
Deontological (D)	Anderson et al. 2004 (W.D.) [11], Anderson et al. 2006 [12], Anderson et al. 2008 [8], Anderson et al. 2014 [10], Bringsjord et al. 2012 [74], Dennis et al. 2016 [139], Malle et al. 2017 [324], McLaren 2003 [336], Mermet et al. 2016 [340], Neto et al. 2011 [359], Noothigattu et al. 2018 [363] Reed et al. 2016 [411], Shim et al. 2017 [455], Turilli 2007 [496], Wiegel et al. 2009 [522]
Consequentialist (C)	Abel et al. 2016 [2], Anderson et al. 2004 (Jeremy) [11], Armstrong 2015 [23], Cloos 2005 [106], Dennis et al. 2015 [140], Dang et al. 2017 [500], Vanderselst et al. 2018 [502], Winfield et al. 2014 [524], Atkinson et al. 2008 [31]
Particularism (P)	Ashley et al. 1994 [28], Guarini 2006 [199]
Hybrid dominance D-C	Arkin 2007 [18], Azad-Manjiri 2014 [35], Dehghani et al. 2008 [134], Govindarajulu et al. 2017 [194], Pereira et al. 2007 [380], Tufis et al. 2015 [495]
Hybrid dominance C-D	Pontier et al. 2012 [393]
Hybrid undefined dominance	Lindner et al. 2017 [307] *(C & A)*, Yilmaz et al. 2017 [540] *(D, C & A)*, Honarvar et al. 2009 [240] *(C & P)*, Howard et al. 2017 [253] *(P & Virtue ethics)*, Berreby et al. 2017 [60] *(D & C)*
Configurable ethics	Bonnemains et al. 2018 [67], Cointe et al. 2016 [110], Ganascia 2007 [185], Thornton et al. 2017 [484]
Ambiguous (A)	Han et al. 2012 [210], Cervantes et al. 2016 [91], Madl et al. 2015 [320], Verheij et al. 2016 [505], Wallach et al. 2010 [515], Wu et al. 2017 [532], Arkoudas et al. 2005 [22], Furbach et al. 2014 [182]

About a quarter of the approaches are of a hybrid nature, combining at least two classical ethical theory types. Approximately half of those have a hierarchical approach, in which deontological features are standardly dominant over consequentialist ones. The non-hierarchical systems, where at least two ethical theory types work together without a single one being dominant, frequently go beyond the two main types of theory. Examples are virtue ethics and particularism [253], and a reflective equilibrium approach that combines consequences, rules, and other influences [540].

A little less than 10% of the papers do not have a specific theory implemented. Instead, they provide various proposals on how to implement different ethical theory types without choosing a particular one. This can be considered a computer scientist approach, where the goal is to devise a general framework which the users can adapt to their preferences.

It is surprising that despite previous calls that a single classical theory is not enough to create an ethical machine and hybrid methods are needed (e.g., [7]), there is relatively little work on hybrid ethical machines. While most hybrid systems have emerged over the last ten to fifteen year, we could not find evidence for an increase in the creation of such systems.

2.7.2 Implementation Classification

Table 2.6 provides an overview of the classification of the non-technical implementation.

Table 2.6: Non-technical dimension classification. Diversity consideration: ✓ implies yes, an empty cell implies no/not present.

Appr.	Contribution type	Eval. type	Eval. subtype	Diversity	Domain	Papers
Top-down	Model representation	Proof	Model checker	✓		Ganascia 2007 [185]
			Logical proof			Arkoudas et al. 2005 [22] Bringsjord et al. 2012 [74] Govindarajulu et al. 2017 [194]
		Informal	Example scenario(s)			Berreby et al. 2017 [60]
		None	None	✓		Bonnemains et al. 2018 [67]
	Model selection	Informal	Example scenario(s)	✓		Turilli 2007 [496]
				✓		Verheij et al. 2016 [505]
				✓		Wiegel et al. 2009 [522]
	Judgment provision	Test	Expert		Engineering	McLaren 2003 [336] Ashley et al. 1994 [28]
			Expert + Non-expert		Military	Reed et al. 2016 [411]
		Proof	Model checker			Dennis et al. 2015 [140]
			Logical proof	✓		Mermet et al. 2016 [340]
		None	None	✓		Lindner et al. 2017 [307]
	Action selection/ execution	Test	Non-expert		Medical	Shim et al. 2017 [455] Dehghani et al. 2008 [134]
			Laws	✓	Cars	Thornton et al. 2016 [484] Vanderelst et al. 2018 [502] Winfield et al. 2014 [524]
		Informal	Example scenario(s)	✓	Medical	Cervantes et al. 2016 [91] Anderson et al. 2008 [8] Cointe et al. 2016 [110]
			Face validity	✓		Pereira et al. 2007 [380]
		None	None	✓	Home care Home care	Neto et al. 2011 [359] Cloos 2005 [106] Dang et al. 2017 [500] Anderson et al. 2004 (Jeremy) [11]
	Judgment provision + action selection/execution	Proof	Model checker	✓		Dennis et al. 2016 [139]
	Model representation + action selection/execution	Informal	Face validity	✓		Atkinson et al. 2008 [31]
Bottom-up	Model representation	Proof	Logical proof	✓		Armstrong 2015 [23]
				✓		Furbach et al. 2014 [182]
	Model selection	None	None	✓		Howard et al. 2017 [253]
				✓		Malle et al. 2017 [324]
	Action selection/ execution	Test	Non-expert	✓		Wu et al. 2017 [532]
		Informal	Example scenario(s)			Abel et al. 2016 [2]
	Model representation + action selection/execution	Test + proof	Non-expert + logical proof	✓	Cars	Noothigattu et al. 2018 [363]
Hybrid	Model representation	Test	Non-expert			Guarini 2006 [199]
			Expert			Anderson et al. 2014 [10]
		None	None		Medical	Azad-Manjiri 2014 [35]
	Action selection/ execution	Test	Non-expert			Honarvar et al. 2009 [240]
			Expert		Medical	Anderson et al. 2006 [12]
			Laws	✓	Medical	Madl et al. 2015 [320]
		Informal	Example scenarios	✓		Yilmaz et al. 2017 [540]
					Military	Arkin 2007 [18]
			Face validity			Han et al. 2012 [210]
		None	None	✓		Anderson et al. 2004 (WD) [11]
				✓		Wallach et al. 2010 [515]
	Model selection + action selection/execution	Informal	Example scenario(s)	✓		Tufis et al. 2015 [495]
	Judgment provision + action selection/execution	Test	Expert	✓	Medical	Pontier et al. 2012 [393]

Approximately consistent with the number of single theory and hybrid theory approaches identified in Section 8.1, most authors choose a top-down approach. Hybrid approaches account for a little less than 25% of those chosen (see Figure 2.2a).

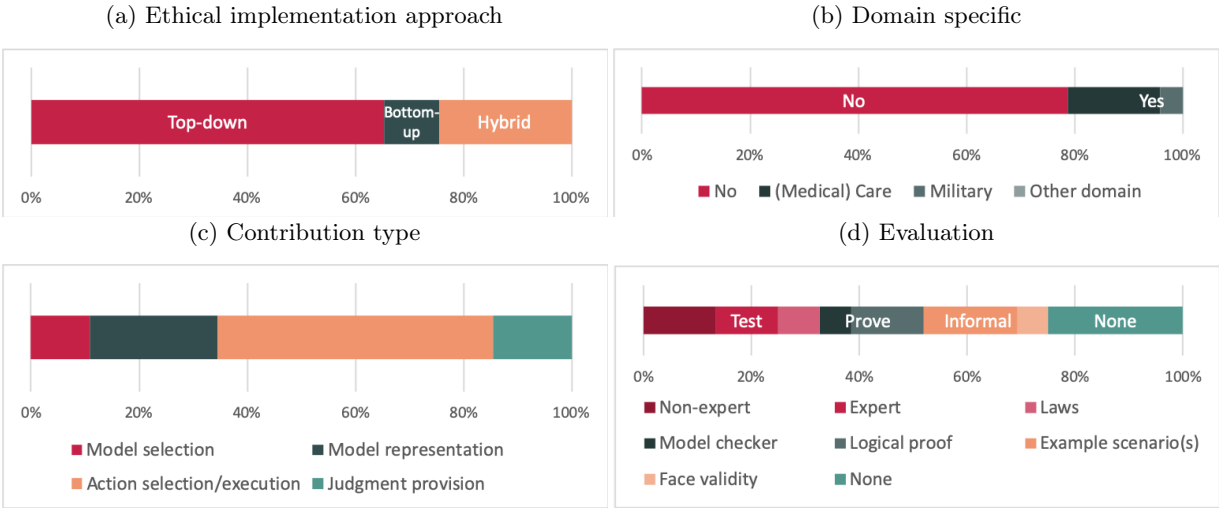


Fig. 2.2: Non-technical analysis

Most authors use a general approach to machine ethics: almost three out of four do not use a domain-specific approach, but focus on a general proposal of implementing machine ethics (see Figure 2.2b).

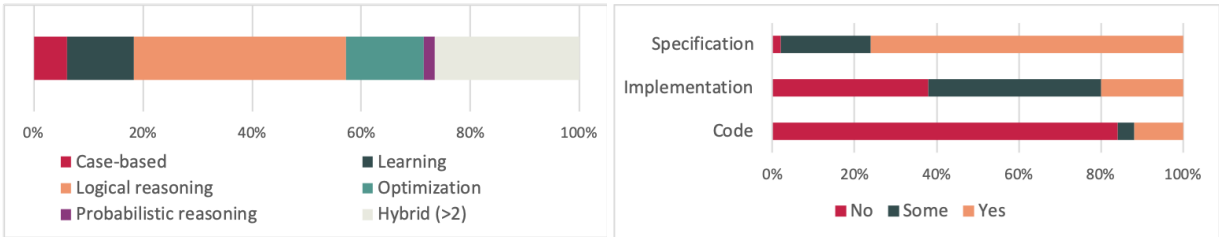
In terms of contribution type, there is a relatively balanced division between authors investigating how an ethical machine should be shaped (model selection and model representation) and authors focusing on the output of the ethical machine (action judgment and action selection/execution, see Figure 2.2c). Most papers address action selection/execution. About 15% of all the papers focus on action judgment: the system judges a situation but leaves it up to the human to actually act on this. From a broader scientific perspective, it is good that both model shaping and output-oriented contributions are investigated. However, it would be ideal to have both things connected.

A possibility for future improvement regards system evaluation: over half of the authors either provide no or only an informal evaluation of their system. Of the rest, about 50% use a test approach and 50% validate their claims with some form of formal proof (see Figure 2.2d).

Finally, about half of the selected papers (51%) acknowledge diversity in implementable ethics, while the other half presents work allowing for or assuming only one ethical theory type.

2.7.3 Technical Classification

The technical dimension classification can be found in Table 2.7. Of the different techniques, logical reasoning is the most frequent. Figure 2.3a shows the distribution of types of technology used. About a quarter of the papers adopt more than one technology type. Only about 10% of the authors focused on a pure learning approach. Case-based reasoning and probabilistic reasoning are the least popular. Mostly classical AI approaches are used—perhaps due to the direct correspondence of rules with deontological ethics.



(a) Technology type

(b) Provided implementation details

Fig. 2.3: Technology analysis

The level of implementation detail provided is somewhat limited (see Figure 2.3b): although most authors include a specification of their idea in the paper, implementation details (or even source code) are rarely included. Both from a computer science perspective and a general science perspective, this is quite undesirable, as it hampers the reproducibility and extensibility of systems and empirical studies.

The different types of input used are fairly distributed: in about 36% of the ethical machines the input is defined as logical cases, in 21% the input has a numerical representation, in 30% the input is written in (natural or structured) language, and 34% use (simulated) sensor data as input. Of all cases, five selected papers had more than one type of input for their system. Around 25% of the authors used a (simulated) robot, corresponding with the amount of sensor data used as input.

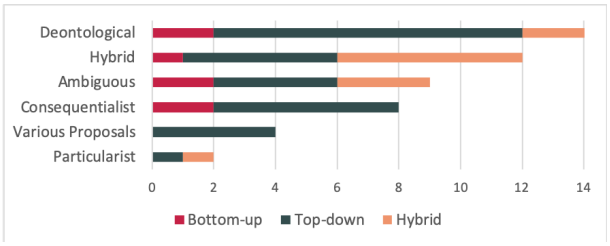
In terms of user friendliness, the implemented systems score poorly. While it is important to note many of these machines are in their prototype phase and more focused on the ethics than the user, it should be important to keep the user in mind from the start of development. Nearly 35% of the machines provide an explanation of their output. 27% process the input automatically, implying that about three out of four implementations require the user to pre-process the input manually in some way — which does not make it easy for the user. Only around one out of five machines include a user interface and less than 17% offer the option for the user to give feedback. In summary, there is still plenty of room for improvement as regards user friendliness.

Table 2.7: Technical classification. ✓ implies yes/fully, ○ implies partially, an empty cell implies no/not present.

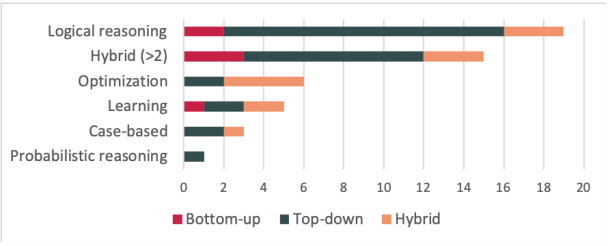
Tech type	Tech subtype	Input				Availability			Other					Papers
		Case (logical rep)	Case (numerical rep)	Case ((struc) language rep)	(Simulated) Sensor data	Formalization details	Implementation details	Code (link) provided	Robot (Simulation)	Feedback	Explanation	UI	Automated Processing	
Logical reasoning (LR)	Deductive logic	✓				✓	○				○			Bringsjord et al. 2012 [74]
		✓				✓	○							Mermet et al. 2016 [340]
		✓				✓	○							Verheij et al. 2016 [505]
	Non-monotonic logic (N-M logic)	✓				✓	✓	✓			○			Ganascia 2007 [185]
	Deontic logic (Deon Logic)	✓				✓	○				○	○		Arkoudas et al. 2005 [22]
		✓				✓					○			Furbach et al. 2014 [182]
		✓				✓			○	○			✓	Malle et al. 2017 [324]
		✓				✓	○							Wiegel et al. 2009 [522]
	Rule-based system (Rules)			✓		✓	○		✓					Dennis et al. 2015 [140]
				✓		✓	○							Dennis et al. 2016 [139]
				✓		✓	○		✓					Neto et al. 2011 [359]
			✓			✓								Pontier et al. 2012 [393]
		✓		✓		✓	○							Tufis et al. 2015 [495]
													✓	Turilli 2007 [496]
	Event calculus	✓				✓								Bonnemains et al. 2018 [67]
	Abductive logic	✓				✓	✓							Pereira et al. 2007 [380]
	N-M logic + event calculus	✓				✓	✓	○			○			Berreby et al. 2017 [60]
	Rules + KR & ontologies			✓		✓	✓	✓						Cointe et al. 2016 [110]
	Deon logic + event calculus	✓				✓	✓	✓						Govindarajulu et al. 2017 [194]
Probabilistic reasoning (PR)	Bayes' Rule + Markov models			✓		○			✓				✓	Cloos 2005 [106]
Learning (L)	Reinforcement learning		✓	✓		✓		✓			○			Abel et al. 2016 [2]
					✓	✓	○						✓	Wu et al. 2017 [532]
	Neural networks		✓			○	✓							Guarini 2006 [199]
			✓		✓	✓	✓			✓				Honarvar et al. 2009 [240]
	NN + Evolutionary computing		✓			○				○				Howard et al. 2017 [253]
Optimization (O)	Optimization			✓		○	○				✓	✓		Anderson et al. 2004 (Jeremy) [11]
				✓		○				✓	✓	✓		Anderson et al. 2004 (WD) [11]
			✓	✓		✓								Anderson et al. 2008 [8]
					✓	✓	○		✓				✓	Thornton al. 2017 [484]
				✓		○	○					✓		Dang et al. 2017 [500]
					✓	✓			✓				✓	Vanderelst et al. 2018 [502]
Case-based reasoning	Case-based reasoning	✓		✓		✓					✓			Atkinson et al. 2008 [31]
		✓		✓		○	○				✓			Ashley et al. 1994 [28]
				✓		✓	✓				✓		✓	McLaren 2003 [336]
LR + L	Inductive logic		✓	✓		✓	✓	✓			✓	✓		Anderson et al. 2014 [10]
	KR & ontologies + inductive logic			✓		○	○				✓	✓		Anderson et al. 2006 [12]
LR + O	Deductive logic + O			✓		✓	○			✓		✓		Yilmaz et al. 2017 [540]
	Rules + O				✓	✓	○		✓	✓	✓	✓	✓	Arkin 2007 [18]
				✓	✓	✓		○	✓					Cervantes et al. 2016 [91]
			✓			✓								Reed et al. 2016 [411]
					✓	✓	○		✓		✓	✓	✓	Shim et al. 2017 [455]
					✓	✓	○		✓				✓	Winfield et al. 2014 [524]
	Rules + abductive logic + O	✓				✓	✓				○			Han et al. 2012 [210]
LR + PR	Rules + Bayes' Rule	✓				✓	○	✓	○			○		Lindner et al. 2017 [307]
	Rules + statistical inference				✓	○	○		✓			✓	✓	Madl et al. 2015 [320]
					✓	○			✓	✓			✓	Wallach et al. 2010 [515]
LR + CBR	Rules + KR & ontology + CBR			✓		○	○						○	Dehghani et al. 2008 [134]
LR + L + O	Rules + decision tree + O		✓			✓					✓			Azad-Manjiri 2014 [35]
PR + O	Bayes' Rule + O			✓		✓				✓				Armstrong 2015 [23]
L + O	Inductive logic + O		✓			✓	○							Noothigattu et al. 2018 [363]

2.7.4 Interactions Between Dimensions

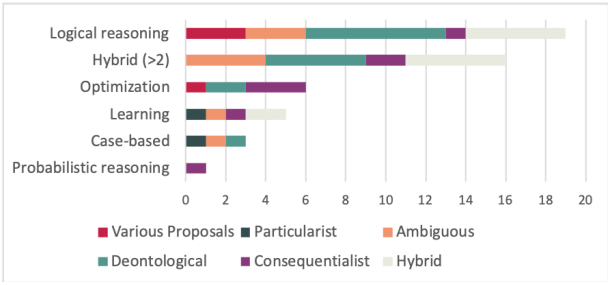
Given that machine ethics is an interdisciplinary field, it is interesting to look at the interaction between the ethical theory types and their implementation, Figure 2.4 shows the interactions between ethical theory, ethical implementation approach, and technology type used to implement an ethical machine. For researchers, it can be useful to see which combinations have not yet been tried out that might be promising. For example, Figure 2.4a shows that (to the best of our knowledge) a hybrid approach (including both top-down and bottom-up elements) to implementing pure consequentialism does not yet exist. Similarly, bottom-up approaches to optimization (see Figure 2.4b) or pure deontological approaches to learning (see Figure 2.4c) (e.g., seeing which input leads to behavior adherent to a certain set of rules) have not yet been explored.



(a) Ethical theory type versus non-technical aspects



(b) Non-technical versus technical aspects



(c) Ethical theory type versus technical aspects

Fig. 2.4: Dimension interaction

2.7.5 General Observations

There are some general observations to be made about the field. Firstly, the focus is on one universal and objective moral agent. There are barely any options for adding cultural influences or societal preferences in any of the classified papers. Almost all systems assume the user cannot influence the output of the system. A recent publication shows indication of cultural differences in ethical preferences [34], and the development of societal preferences within an ethical machine would improve the chance of acceptance of ethical

machines. However, it is still under debate whether the field should move towards a "universal moral grammar," such as that proposed by Mikhail [342].

Secondly, there are some issues inherent to the field. For instance, there are no benchmarks to verify if a system is working as it should. There are no specific tasks to be implemented, no consensus as to what the correct output is, and few data sets to use in an implementation. A helpful tool to recur to in this context is the work by Whitley [521], who provides a four-dimensional schema for analyzing a research field. Two of the dimensions refer to the uncertainty of the task at hand, and two refer to the mutual dependence between the fields and scientists in them. The field of machine ethics scores highly on all of these dimensions:

High technical task uncertainty: there is unpredictability and variability in which methods are used in the field and how results are interpreted. In this regard, it is a fragmented field.
High strategic task uncertainty: there are problems present in the field that are valued differently (e.g., some authors focus on the theoretical, others on the implementation, and the ethical theories or even ethical theory types they focus on diverge).
High strategic dependence: there is much disagreement on the relevant topics, so there is a high reliance on peers for validation and reputation in the field.
Medium functional dependence: in terms of physical dependence of resources, there is none. Anyone with a computer can add to the field; no expensive equipment is needed. However, there is a high dependence on results of others and acceptance by the field.

Another potentially helpful perspective can be derived from Whitley's theory, where the field of machine ethics would be a "polycentric oligarchy", implying there are several independent clusters of scholars that confirm each other's assumptions and do not communicate much with other clusters that have very different views. At first glance, such clusters can indeed be detected: the multi-agent norm domain (e.g., [359, 495]), the logical translation of ethical theories (e.g., [194, 210, 380]) or the modern learning approach to machine ethics (e.g., [2, 532]). While exploratory research in many directions is valuable, the field would benefit from more standardization and more communication between clusters to exchange knowledge on ethics and technology.

2.8 Future avenues and limitations

Based on the results of the analysis and description of the selected papers, some literature gaps are identified that can be of interest for future work. Additionally, the limitations of this survey are discussed.

Ethical dimension In view of earlier calls for hybrid systems when it comes to ethical theory, a surprisingly low percentage of authors consider a **multi-theory** approach in which machines can interchangeably apply different theories depending on the type of

situation. In terms of the content (and not the structure) of ethical theories, it is important to acknowledge and harness the nuances of specific theories, but human morality is complex and cannot be captured by one single classical ethical theory. Even experts can have rational disagreement amongst themselves on an ethical dilemma. This leads to the next important point: an ethical machine will not be of use if it is not accepted by its users, which can be the risk of focusing on one ethical theory and, thus, not covering human morality. Ethical theory needs to be combined with **domain-specific ethics** as accepted by domain experts and, as identified in the analysis of this paper, this is not the case in the majority of the related work. Moreover, it is necessary to discuss the ethical theory/theories in the system with its possible users. Some examples of using **folk morality** in machine ethics can be found in Noothigattu et al. [363], as well as in [422]. However, it is important to note that just as ethical theories have their challenges, so does folk morality. Three challenges are who to include in the group whose values should be considered (*standing*), how to obtain their values (*measurement*), and how to aggregate their values (*aggregation*) [47]. Implementations should start from ethical theories combined with domain-specific ethical theory, after which acceptance by the users and deviation from socially accepted norms should be discussed (cf. e.g., [34, 66, 301, 440]).

Non-technical dimension There is a need for more systematic **evaluations** when ethical machines are created in order to be able to rate and compare systems. To this end, there is a strong need for **domain-specific benchmarks**. Based on input from domain experts, data sets need to be created containing the types of cases prevalent in that domain, with respect to which ethical machines must be assessed. The gathering of typical tasks and respective answers that domain experts agree on is just as important as the actual creation of ethical machines. This implies the need for more **collaboration** between fields. Computer scientists and philosophers, as well as domain experts and social science experts, have to work together to ensure the interaction with and effects of the ethical machines are as desired. Even within the field, collaboration is needed between different clusters of topics in the field of machine ethics, for example between clusters specializing in MAS and machine learning respectively. Finally, in general, **implementation** requires more attention. While on a higher level, theoretical discussion remains important in this field, especially to prepare for possible future scenarios, the testing of theory in practice can enrich the discussion on what is (or is not) possible at that moment and what practical implementations and consequences certain ethical machines can have.

Technical dimension When a system is implemented, it is imperative to provide exhaustive specification detail, including **availability of the code**, which is predominantly lacking. Another frequent shortcoming regards usability: the system should have a user interface so that the future user can interact with the system without having to know how to code. Furthermore, automatic processing of input cases deserves more attention, so as to avoid having to encode each variable manually as a vector for a neural network. Considering the

increased need for **transparency** in algorithmic decision making, as well as the fundamental role of reasons in ethics, the system should also provide an explanation of why it took a certain decision. In a next phase, the user should be able to give **feedback** on the ethical decision the system makes. Finally, the association of a given type of technology with a certain type of ethics requires an adequate technical justification, beyond using just the most acquainted technology.

Further Points of Interest Current technology allows for successful application of narrow AI geared towards specific tasks. While steps are being taken towards AGI, the technology does not yet exist [263]. Hence, domain-specific applications seem suitable. A domain-specific non-AGI approach to machine ethics alleviates some of the risks and limitations on machine ethics posed by [77], such as those related to an "insufficient knowledge and/or computational resources for the situation at hand." However, there are still risks and limitations. For instance, in the context of lethal autonomous weapons systems, the loss of "meaningful human control" [430] is a risk, as humans would not have the same control over ethical decisions such as target selection. A limitation of using domain-specific ethical machines is that the process of one domain may not be transferable to other domains. Furthermore, not everyone is ready to accept a machine taking over the ethical decision making process [228].

A slightly different way to address ethics in machines is to define (and implement) an ethical decision support, rather than leaving the machine to make an autonomous ethical decision. For an overview of different types of moral mediation, see Van de Voort et al. [501]. Etzioni agrees that the focus should lie on decision support, stating "there seem to be very strong reasons to treat smart machines as partners, rather than as commanding a mind that allows them to function on their own" [158, p. 412]. One of those reasons is that AGI will not exist in the foreseeable future. This approach will also help with acceptance of machines with ethical considerations in society. There are different possible levels of autonomy the system can have, for example only summarizing available data, interpreting available data, summarizing possible actions, or even suggesting/pre-selecting a possible action the system deems best. Different types of support and collaboration might be necessary for different applications, and according to the literature review done in this paper, further research is needed in this direction.

Limitations This survey has some limitations that need to be mentioned. First of all, the scope of the paper selection was limited to explicit ethical theories (i.e., theories directly programmed into the machine). While some of the works reviewed can still be of interest and provide inspiration for implementation, papers devoid of implementation details were excluded from this survey. Examples are emerging ethics based on human data to research folk morality (e.g., [532]) or models of human morality to determine relevant features in input cases (e.g., [511]). Furthermore, we limited the survey to one paper per author whenever similar systems were discussed across multiple publications, selecting

the most comprehensive one. This does not do full justice to the work of certain authors (e.g., Guarini working on explainability of neural networks making ethical decision [200]). While the paper selection procedure was designed to be as exhaustive as possible, it is still possible that a few important papers were missed. Finally, three authors reviewed the ethical dimension and two reviewed the implementation and technical dimension, but it is still possible there was bias in the classification due to the limited number of people involved in the classification process and the process of discussion until agreement was reached.

2.9 Conclusion

The future of the field of machine ethics will depend on advances in both technology and ethical theory. Until new breakthroughs change the field, it is important to acknowledge what has been done so far and the avenues of research that make sense to pursue in the near future. To accomplish this, the contribution of this survey is threefold. Firstly, a classification taxonomy with three dimensions is introduced: the ethical dimension, the dimension considering nontechnical aspects when implementing ethics into a machine, and the technical dimension. Secondly, an exhaustive selection of papers describing machine ethics implementations is presented, summarized, and classified according to the introduced taxonomies. Finally, based on the classification, a trend analysis is presented that leads to some recommendations on future research foci. It is important to keep in mind how machine ethics can be used in a meaningful way for its users, with increasing agreement on what a system should do, and in what context.

2.10 Acknowledgement

This work is partially funded by armasuisse Science and Technology (S+T), via the Swiss Center for Drones and Robotics of the Department of Defense, Civil Protection and Sport (DDPS).

We thank the authors of the selected papers for providing valuable feedback on their paper's representation. Also, we would like to thank the anonymous reviewers for their feedback on our manuscript, which has helped us to substantially improve it.

AI Mistakes and Trust Formation

This chapter is based on:

Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. ***Second Chance for a First Impression? Trust Development in Intelligent System Interaction****. In Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21). Association for Computing Machinery, New York, NY, USA, 77–87. DOI: https://doi.org/10.1145/3450613.3456817*

and on:

Suzanne Tolmeijer, Astrid Weiss, Marc Hanheide, Felix Lindner, Thomas M. Powers, Clare Dixon, and Myrthe L. Tielman. ***Taxonomy of Trust-Relevant Failures and Mitigation Strategies****. In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction. Association for Computing Machinery, New York, NY, USA, 3–12. DOI: https://doi.org/10.1145/3319502.3374793*

Second Chance for a First Impression? Trust Development in Intelligent System Interaction

Suzanne Tolmeijer[1], Ujwal Gadiraju[2], Ramya Ghantasala[2], Akshit Gupta[2], and Abraham Bernstein[1]

[1] Department of Informatics, University of Zurich, Switzerland
[2] Faculty of Engineering, Mathematics and Computer Science, Delft University of Technology, Netherlands

Abstract. There is a growing use of intelligent systems to support human decision-making across several domains. Trust in intelligent systems, however, is pivotal in shaping their widespread adoption. Little is currently understood about how trust in an intelligent system evolves over time and how it is mediated by the accuracy of the system. We aim to address this knowledge gap by exploring trust formation over time and its relation to system accuracy. To that end, we built an intelligent house recommendation system and carried out a longitudinal study consisting of 201 participants across 3 sessions in a week. In each session, participants were tasked with finding housing that fit a given set of constraints using a conventional web interface that reflected a typical housing search website. Participants could choose to use an intelligent decision support system to help them find the right house. Depending on the group, participants received a variation of accurate or inaccurate advice from the intelligent system throughout each session. We measured trust using a *trust in automation* scale at the end of each session.
We found evidence suggesting that trust development is a slow process that evolves over multiple sessions, and that first impressions of the intelligent system are highly influential. Our results echo earlier research on trust formation in single session interactions, corroborating that reliability, validity, predictability, and dependability all influence trust formation. We also found that the age of the participants and their affinity with technology had an effect on their trust in the intelligent system. Our findings highlight the importance of first impressions and improvement of system accuracy for trust development. Hence, our study is an important first step in understanding trust development, breakdown of trust, and trust repair over multiple system interactions, informing improved system design.

3.1 Introduction

Technological advances in storage and computation have led to the unprecedented rise in the use of artificial intelligence (AI) and automation. This has resulted in the widespread adoption of intelligent systems across several domains including healthcare, transport, manufacturing, finance, and education [406]. Many everyday tasks are supported by AI systems today. From data-fueled cloud services on computers to smart apps on mobile phones, intelligent decision support is becoming increasingly ubiquitous. Although such support can make life easier for users, inappropriate reliance can also lead to failures [293]. Consider the example of a navigation support system. On the one hand, misuse or absolute reliance on the system can lead a user to follow an outdated speed limit. Disuse or lack of reliance on the system on the other hand, can lead to missed benefits, such as a user getting stuck in traffic due to the lack of trust in a suggested detour. Considering

that AI systems are now penetrating critical domains [367], one can expect far graver consequences of user trust or the lack of it in such systems.

With AI playing a prominent role in our lives, important questions surrounding our trust in AI systems have emerged. How exactly does trust evolve in the interaction between humans and AI systems? To what extent is the trust that is established through interaction robust to system accuracy over time? What factors mediate trust formation? Since trust in intelligent systems is fundamental to their widespread adoption, these are pivotal knowledge gaps to address in the emerging field of *Human-AI Interaction*.

We adopt the following definition of trust as: "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [293]. Trust is relevant when a situation contains a truster and trustee. The trustee has a task to perform with an incentive to perform it and the truster has the uncertainty and risk of failing the task [214]. Along with dispositional factors such as age and situational factors such as subject expertise, trust is learned over time [233].

To our knowledge, there has been little research on dynamically learned trust that evolves over different interactions with a system [237]. In particular, the influence of accuracy and reliability on trust formation over time have been insufficiently explored. However, learning about trust development is vital for successful system usage over time. Additionally, while dispositional factors such as age and affinity with technology have been shown to influence trust [233, 437], little is understood about their interaction with system accuracy. Thus, we pose the following research questions:

RQ1 *Does the accuracy of advice of an intelligent system over multiple sessions influence the reliance of users on that advice?*

RQ2 *Does inconsistency of accurate advice from an intelligent system over multiple sessions influence trust formation?*

RQ3 *Can inaccurate advice from an intelligent system harm trust formation and accurate advice recover trust formation over multiple sessions?*

RQ4 *Do dispositional factors such as age and propensity to trust influence trust formation in an intelligent system across multiple sessions?*

To investigate these questions, we considered a domain relevant to our everyday lives, and built an intelligent housing recommendation system to carry out a multi-session study consisting of 201 participants across 3 sessions in a week. In each session, participants were tasked with finding houses that fit a given set of constraints using a housing search website that we created (as shown in Figure 3.1). Participants could choose to use an intelligent decision support system to help them find the right house. The tasks were designed to make manual search relatively taxing, encouraging participants to use the intelligent system. We offered a return bonus to increase the chance of participants returning for all sessions

as well as a task bonus to incentivize finding the correct task solution. Depending on the group, participants received a variation of accurate or inaccurate advice from the intelligent system throughout each of the sessions. We measured trust in the system at the end of each session using the established 'trust in automation scale' [269].

Original Contributions. In this work, we present experimental evidence which suggests that first impressions matter for trust formation in Human-AI interaction over multiple sessions. However, trust can be recovered and even improved significantly when a faulty first session is followed by consistent and accurate user support by an intelligent system. Trust formation shows slow upward and downward trends, confirming that trust develops over time and is influenced by system predictability and reliability. Finally, we find that the age of the user and their affinity with technology correlate with trust development. Our findings inform system designers of the importance of first impressions and (appearance of) system improvement over time during multiple interactions. Our contributions through this work inform future research directions pertaining to trust formation, loss of trust, and trust repair. We publicly share all our data, to promote open science.[3]

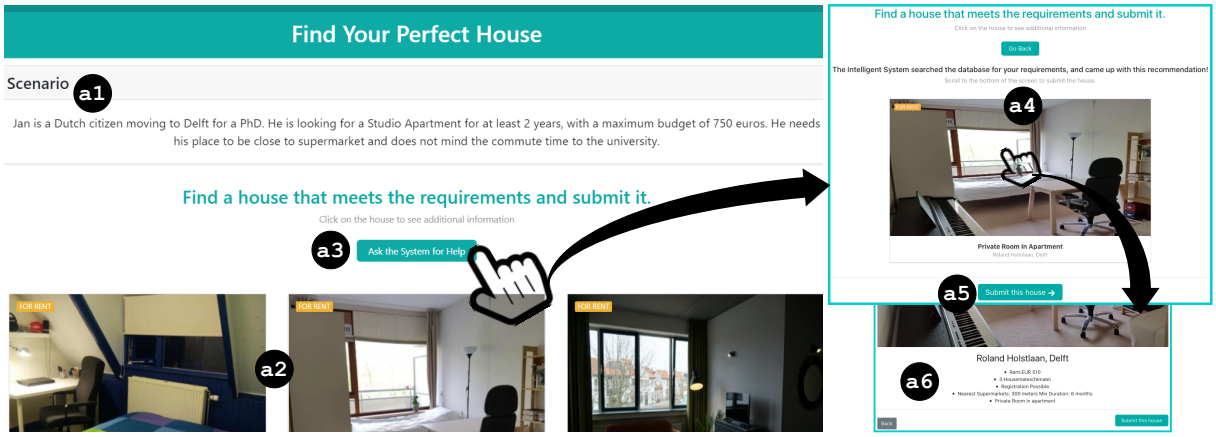


Fig. 3.1: The housing search interface (left-hand side), and assistance from the intelligent system (right-hand side).

3.2 Related work

We discuss related literature in three realms: (i) how trust has been modeled and studied in HCI, (ii) trust formation in user interactions with intelligent systems, and (iii) the relationship between user trust and system accuracy.

[3] Open Science Foundation (OSF): https://osf.io/ndjfs/?view_only=502f2abc34714838918213a04d68dc58

3.2.1 Trust in Human-Computer Interaction

The interest of the HCI community in trust is apparent in recent literature. From trust in automation [293] and intelligent systems [157, 237] to trust in AI, machine learning, and robotics [460], prior works have explored trust in various systems over the years. Hoff and Bashir [233] have integrated research on trust factors into an overall model. According to them, trust in automation has three main components: dispositional trust, situational trust, and learned trust. Our focus lies on learned trust, which consists of initial learned trust (including expectations) and dynamic learned trust (which changes during the interactions with the system). Hoff and Bashir identify a research gap on how previous positive and negative experiences and resulting expectations influence trust in future interactions, which is the focus of our study.

Schaeffer et al. [437] also focus on factors impacting trust formation in automation in their meta-analysis. The four main influence categories of their model include 1) traits such as age, 2) emotive factors such as attitudes towards the system, 3) states including stress, and 4) cognitive factors such as expectancy. Among other research gaps, they list a lack of research on age impact, reliability and errors – all of which are discussed in this study.

3.2.2 Trust Formation

Trust develops over time and depends on many factors. Each interaction with a system alters the trust in that system. Holliday et al. [237] looked at trust formation within one user session. They found that the impression of a learning system, conveyed through explanations, led to higher levels of trust. In addition to a system learning over time, the impression of system reliability shapes trust. Case in point, consistent reliable support leads to steadily increasing trust, while consistent unreliable support led to constant decrease in trust [50]. First impressions are especially important: negative first impressions have a stronger negative influence on trust than negative impressions acquired later on [366].

Understanding trust formation does not only involve how trust is fostered, but also when it breaks and how it can be recovered. Trust break and recovery have been understudied [130, 492]. In this study, we thereby focus on the influence of accuracy on trust formation and whether improved accuracy is enough to regain trust after inaccurate advice.

3.2.3 Trust and Accuracy

The influence of accuracy on trust has become more influential as artificial intelligent methods have become more opaque, e.g., when compared to earlier rule-based system. While results from AI have been very promising, users do not trust what they do no understand [365]. In fact, providing explanations for AI models that are less human-meaningful

decreases perceived accuracy compared to actual accuracy [365]. The importance of the impression of the system is echoed in work by Yin et al [541]. Authors found a difference in trust formation between the effect of stated accuracy and observed accuracy: stated accuracy has a significant effect on trust independent of actual accuracy. Nevertheless, model accuracy is more important for trust formation than explanations [374].

If the system is indeed unreliable or inaccurate, the user takes longer to decide whether to follow the system's advice [489]. In robots, Desai et al. [141] found that early unreliability had a greater impact on trust formation than unreliability later on. Additionally, the error type also determines the impact on trust formation. For instance, in the autonomous cars domain, obstacles that were not detected but missed had a bigger impact than false alarms of obstacles [36].

A study on accuracy over time with multiple sessions was done [242], but in relation to user feedback. They found that allowing users to provide feedback lowered trust in the system and lead to a lower experienced accuracy, independent of actual system accuracy. To our knowledge however, an in-depth understanding of the interaction between accuracy and trust formation over time is missing - especially whether (in)accuracy can lead to trust loss and trust recovery.

3.3 Study Design

To address the aforementioned research questions, we conducted a crowdsourced multi-session study. In this section, we describe the measures, task design, and the procedure.

3.3.1 Measures

Measuring User Trust in the Intelligent System. We used a validated trust scale [269] to measure user trust in each case. The scale consists of 12 items pertaining to the intelligent system, and participants are asked to use a 7 point Likert-scale ranging from (*1: Not at all*) to (*7: Extremely*), to indicate their agreement with each item. While relatively recent scales for trust measurement such as the multi-dimensional measure of trust are available [326] or domain specific trust scales such as for online recommender agents have been proposed [51], we chose to use a more generic and validated scale of trust in automated systems [269]. To account for the dispositional component of user trust formation, we additionally used the validated and widely accepted 'propensity to trust scale' [179]. Each trust scale was aggregated into an average trust score per participant ranging from 1-7 and 1-5 respectively. In the case of the trust in automation scale, scores of negatively worded trust were reverse coded.

Measuring Affinity for Technology. Recent research has shown that affinity for technology interaction can be seen as a core personal resource for successful coping with technology and a facet of user personality [33]. We used the 9-item 'affinity for technology interaction' (ATI) to assess a user's tendency to actively engage in technology interaction [176].

3.3.2 Task Design and the Intelligent System

Trust requires three components: actors to form trust, an incentive to trust, and a risk to trust [214]. We modeled our task to integrate these three components. In the task, participants (*the trusters*) were presented with house searching scenarios with a given set of constraints. There was only one house per scenario that fit all constraints. Finding the right house that satisfies all requirements was rewarded with a monetary bonus of 0.25 GBP (*the incentive*). Participants were offered advice by an intelligent system (*the trustee*). If they did not consider the advice, they risked *losing valuable time* by having to manually click through each of the displayed houses to find the right one that matched all constraints (*the risk*). We considered the housing domain since many people have experience with it and items naturally need to fit multiple requirements, making the search challenging enough to benefit from automated assistance.

Figure 3.1 presents an overview of the interface and the intelligent system that users were equipped with. On beginning the task, a house search scenario is presented to the user at the top of the interface (cf. (a1)). The scenario describes the constraints pertaining to the house search, in a situated search format. We manually crafted the tasks to be taxing, to create a realistic incentive for the users to engage with the intelligent system. We considered two levels of complexity within the house search task: in the relatively **`easy`** scenario, users were assigned a house search task with 3 constraints, while they had to deal with 5 constraints in the **`complex`** scenario (as shown in Table 3.1).

Table 3.1: Examples of `easy` and `complex` scenarios presented to users in each house search task. Each distinct constraint is colored for the benefit of the reader.

Complexity	Scenario
`Easy`	Peter is moving to Delft as a first year Bsc. student. He is a very easy going guy and is looking for a shared room which fits his rent budget of 300€. Further, he would require registration at the municipality.
`Complex`	Jan is a Dutch citizen moving to Delft for a PhD. He is looking for a studio apartment for at least 2 years, with a maximum budget of 750€. He needs his place to be close to a supermarket and does not mind the commute time to the university.

Note that there was a total of 12 houses displayed on the interface in a randomized order (cf. (a2)), and in each task only one house satisfied the given constraints. Participants could use assistance from the intelligent system by clicking the **Ask the System for Help** button, present below the scenario description (cf. (a3)). On clicking the button, the intelligent system presents the user with a house, that it claims matches all the required scenario constraints (cf. (a4)). Users can either submit the house directly by using the **Submit this House** button (cf. (a5)), or verify whether the constraints corresponding to the suggested house are indeed satisfied, by clicking on it and viewing the details (cf. (a6)). Based on the experimental condition (described in the following section), the intelligent system either provided an accurate or inaccurate suggestion. Users could freely switch between manually sifting through each house and using the intelligent system by using the **Back** button. By clicking on a house, users could view its details.

System Implementation. We created a web application using React.js for the front-end of the house search interface and Node.js as well as Express.js for the backend. MongoDB was used both for logging user interactions in the task and for storing data pertaining to the houses. The application was hosted on Heroku. In total, we created six distinct scenarios: three **`easy`** and three **`complex`**. Each participant was then randomly assigned one easy scenario and one complex scenario in each of the three sessions. The scenarios were also randomized across the sessions and between groups for participants, to prevent biases due to ordering effects. The total number of houses in the data set was 12 and for each scenario the position of the correct house in the displayed list was randomized to prevent biases due to ordering or learning effects. A fixed list of incorrect houses was created to support the sessions with incorrect advice from the intelligent system. In such sessions, a random house was selected and displayed from this list as a suggestion from the system. In case of session with correct advice from the intelligent system, the correct house was shown.

Although there are more elaborate systems for housing search (e.g., [435]), we opted for a simpler interface that allowed us to isolate the effect of the system advice on user trust.

3.3.3 Procedure and Experimental Setup

We recruited participants from crowdsourcing platform *Prolific*.[4] The platform has been shown to be an effective and reliable choice for running relatively complex and time-consuming interactive information retrieval studies [419, 535]. Crowdworkers on Prolific were invited to participate in a multi-session study titled, "*Finding the right house that meets your requirements*". To ensure reliable participation, we followed Prolific's guidelines and restricted eligibility to workers who had an acceptance rate of at least 80% and had

[4] https://www.prolific.co

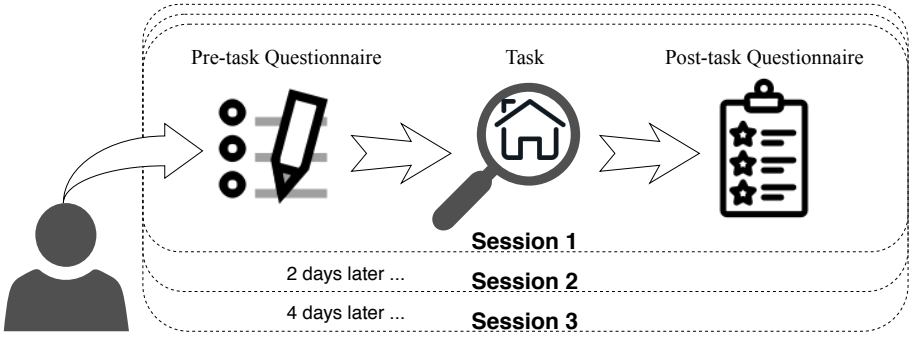


Fig. 3.2: Overview of the study workflow.

at least 10 successful submissions on the platform. Participants were informed about the longitudinal nature of the task. Those who accepted our task received brief instructions about the task and were asked to check-off an informed consent before beginning their task session. As shown in Figure 3.2, participants were first asked to complete a pre-task questionnaire consisting of (i) demographic questions about their *age*, *gender*, and *country of origin*, as well as (ii) *the affinity to technology interaction* (ATI) scale. Next, participants were assigned two consecutive house searching tasks. They were incentivized with a monetary bonus of 0.25 GBP for finding the right houses. On completing the two house search tasks, participants were asked to complete an exit questionnaire consisting of (i) the 'trust scale', (ii) the 'propensity to trust' scale, and (iii) a text area to provide optional remarks or comments. On completing the exit questionnaire, participants received a completion code which they were asked to enter on Prolific to receive their base payment. We paid all participants Prolific's suggested fair wage of 7.50 GBP/h.

After a successful session, participants were invited to join a second session two days later and a third session two days after the second session. To maximize the return rate of participants, we rewarded participants with a return bonus of 1 GBP in addition to their base pay for completing Session 2 and 3. Since we logged participant data using their Prolific IDs, we ensured that in each session participants received two distinct house search scenarios (one **`easy`**, one **`complex`**), which they did not encounter previously. While the study flow was identical in the three sessions, participants were only asked to respond to the demographic questions, and fill out the 'propensity to trust' scale in the first session.

Participants were randomly assigned to one of eight different experimental conditions (referred to as 'groups' hereafter), that differed in the sequence of accuracy of the intelligent system across the three sessions. Assuming that **1** represents accurate advice and **0** represents inaccurate advice given by the system in a given session, the experimental conditions were as follows: '**1 1 1**', '**1 1 0**', '**1 0 1**','**0 1 1**','**0 0 1**', '**0 1 0**','**1 0 0**', and '**0 0 0**'. For instance, this means that participants assigned to

group '**0 1 0**' received incorrect advice in session one, correct advice in session two, and incorrect advice in session three.

3.3.4 Hypotheses

The aforementioned experimental conditions (or groups) allow us to test different hypotheses by comparing sessions and groups. Specifically, the hypotheses we test to answer our research questions can be found in Table 3.2.

Table 3.2: Hypotheses and their required comparisons. Comparisons are made either between sessions within a single group, or between different groups. Investigation of dispositional factors is not related to specific sessions or groups.

RQ	Hypothesis	Comparison	Groups
1	H1. Increased amounts of accurate advice leads to more user reliance, while inaccurate advice will lower intelligent system dependence.	Groups	All groups
2	H2. Consistent inaccurate advice over multiple sessions leads to significantly lower trust than inconsistent accuracy over time.	Groups	'000' vs. '100', '010', and '001'
2	H3. Trust is significantly higher for users that receive consistent accurate advice.	Groups	'011', '101', and '110' vs. '111'
2	H4. Inaccurate advice is more harmful in earlier sessions rather than later sessions.	Sessions	'001' vs. '010' vs. '100' '011' vs. '101' vs. '110'
3	H5. Trust is lost significantly if an inaccurate session follows an accurate session.	Sessions	'110', '010', '101', and '100'
3	H6. Trust does not recover significantly when consistent accurate advice follows an inaccurate first impression.	Sessions	'011'
4	H7. The dispositional factors of gender, age, culture, experience with computer science, propensity to trust, and affinity with technology all influence trust formation across multiple sessions.	Sessions and Groups	All groups

3.4 Results

In our first session, 255 subjects participated. Of those participants, 83% returned for the second session two days later. 96% of these participants returned to complete the third session two additional days later. This resulted in a total of 203 participants, who completed three sessions. Two participants were excluded based on clearly evident unreliable participation. Thus, the results and analysis presented hereafter pertain to these 201 participants unless specified otherwise (see Table 3.3). A compromise power analysis of

the mixed ANOVA revealed that with over 24 participants per group, we have a power of 0.9 (considering a medium effect size of $f = 0.25$, $\alpha = 0.05$).

We found that 26 participants did not use the intelligent system in any of the three sessions. Trust scores for these participants were therefore excluded in our analyses pertaining to user trust. Since questions in the trust scale refer to system performance, the responses from users who never utilized the system are meaningless.

Table 3.3: Number of participants per experimental condition

Group	**000**	**001**	**010**	**011**	**100**	**101**	**110**	**111**
Participants	24	24	26	26	26	24	25	26

Each session consisted of two scenarios. We will refer to sessions using S1 to S3, scenarios will be denoted as S1.1 and S1.2 for each session. To control for Type-I error inflation in our multiple comparisons, we use the Holm-Bonferroni correction for family-wise error rate (FWER) [238], at the significance level of $\alpha < .05$. Significance levels are marked as follows: $p < 0.05$*, $p < 0.01$**, and $p < 0.001$***.

3.4.1 Participant Demographics

One hundred thirty-five participants reported to be female (66 male). The age of the participants ranged from 18 to 62 years old (*M=27.5, SD=9.2*). Education ranged from high school or less (29%) and college without degree (25%) to some form of degree obtained throughout their life (46%). 40% of the participants reported to have studied computer science or some related field. Participants originated from 30 different countries, with most participants reportedly born in the United Kingdom (41), Poland (36), Portugal (26), and Italy (22).

3.4.2 Success of Participants Across Sessions

Independent of the experimental group assignment, on average participants were able to successfully find the right house in 78% of the scenarios in the first session, 66% in the second session, and 92% in the third. Part of the difference in user accuracy in finding the right house can be explained by the difficulty of the scenario: in four out of six scenarios, there is a significant difference, using Fisher's exact test, between **`easy`** and **`complex`** scenarios and correct/incorrect answers given by the participants. Complexity does not explain user accuracy in the second scenario of session 1 and 3. Another explaining variable to user accuracy is correctness of the system's advice. Except for the second scenario of session 1, there is a significant difference between user and system (in)accuracy: users made less mistakes when the system gave correct advice. The summarized results of the two Fisher's exact tests can be found in Table 3.4.

Table 3.4: P-value results of two Fisher's exact tests on user accuracy. Difficulty (`easy/complex`) and system accuracy (correct/incorrect) were compared against user accuracy (correct/incorrect).

Sessions	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2
Difficulty	0.006**	0.854	1.076e-5***	0.011*	0.018*	0.814
System Accuracy	0.001e-1***	0.104	5.634e-6***	1.422e-7***	0.018*	0.003**

3.4.3 System Accuracy influences Reliance

We analyzed the reliance of users on the intelligent system. Indicators of user reliance on the system can be distinguished at two levels: users clicking the **Ask the System for Help** button to open the system's suggestion (*open*) or users submitting the system's suggestion by clicking the **Submit this House** button as their answer (*submit*).

Table 3.5: Ratio of participants who used the system per group by clicking the system suggestion at least once. Average usage ratio per group is shown in the last column.

Group	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2	*avg.*
000	0.54	0.42	0.54	0.42	0.33	0.29	*0.42*
001	0.71	0.50	0.63	0.58	0.63	0.63	*0.61*
010	0.85	0.58	0.65	0.58	0.65	0.62	*0.65*
011	0.54	0.50	0.54	0.58	0.65	0.69	*0.58*
100	0.62	0.69	0.77	0.85	0.58	0.54	*0.67*
101	0.79	0.54	0.71	0.67	0.58	0.58	*0.65*
110	0.68	0.60	0.76	0.84	0.76	0.72	*0.73*
111	0.77	0.73	0.69	0.77	0.73	0.69	*0.73*

Table 3.6: Ratio of participants who submitted the system's suggestion after opening it.

Group	S1.1	S1.2	S2.1	S2.2	S3.1	S3.2	*avg.*
000	0	0.20	0.08	0.20	0.13	0.14	0.12
001	0.12	0	0	0.07	0.67	0.73	0.26
010	0.09	0.07	0.47	0.67	0.06	0.06	0.24
011	0.07	0.31	0.50	0.67	0.53	0.72	0.47
100	0.44	0.72	0.30	0.32	0.07	0.21	0.34
101	0.42	0.46	0.41	0.44	0.71	0.93	0.56
110	0.41	0.67	0.42	0.71	0.11	0.11	0.41
111	0.35	0.63	0.61	0.65	0.84	0.89	0.67

Results pertaining to the reliance of users on the intelligent system can be found in Table 3.5. The group average for *opening* the system was above 50% for all groups except group '**000**'. In this case, usage drops gradually to below 33% in the last session — the only session across all groups where this is observed. Interestingly, system usage within a session sometimes dropped despite the system providing correct suggestions or increased despite inaccurate advice. However, the highest average usage of the system was observed for group '**111**' and lowest for group '**000**'. The order of average usage ratios suggest that first impressions matter, which is further discussed in the next subsection. Nearly all first usage in a session stayed equal or went up if the previous session had correct advice,

or stayed equal/went down if the previous session contained incorrect advice from the system. Exceptions to this trend are the first session transition within group '**101**' and the two session transitions within group '**111**', possibly because system usage was already relatively high to begin with. This partially supports H1: while in general, more accurate advice leads to consulting the intelligent system more often, there are exceptions (such as group **'011'** having less system usage than group **'100'**).

The fraction of users who submitted the system's suggested house after seeing it in each session, and across all groups, is reported in Table 3.6. Of the users who opened the system during a session, the fraction that *submitted* the suggested house ranged from 12% in group '**000**' to 66% in group '**111**' on average. The average submission reliance ratio for all incorrect sessions was 0.15, while it was 0.62 over all correct sessions. Again, we see that first impressions, i.e., correct advise in session 1, influenced reliance on the system, with respect to following the suggestion. This supports H1: once the system is consulted, advice usage increases with system accuracy.

3.4.4 First Impressions Matter

Participants were assigned to different orders of session support accuracy. To answer RQ2, trust scores were compared using a mixed ANOVA between groups within sessions.

We found a statistically significant effect of experimental group assignment on trust scores ($F(7, 146) = 18.456, p < .0001, \eta_p^2 = 0.470$). While session alone did not influence user trust ($F(2, 292) = 0.127, p = 0.881, \eta_p^2 = 0.001$), there was a strong interaction between the accuracy of a group and session in explaining the reported average trust scores ($F(14, 292) = 19.910, p < .0001, \eta_p^2 = 0.488$). Pairwise comparisons between groups are illustrated in Table 3.7, while the pairwise comparisons of sessions within groups are presented in Table 3.8. Tukey's HSD test was used for post-hoc analysis (reported in Table 3.7 and 3.8).

Table 3.7: Results of mixed ANOVA for average trust scores between groups. Green cells imply a significant difference between groups. The group mentioned in a green cell had a higher average trust based on Tukey's HSD test.

Group	000	001	010	011	100	101	110	111
000								
001								
010								
011	011**	011***	011**					
100	100*	100*						
101	101***	101***	101***		101*			
110	110***	110***	110***		110*			
111	111***	111***	111***	111***	111***	111**	111***	

Given that Table 3.7 shows comparisons for aggregated trust scores over all three sessions, we expect groups with equal number of correct suggestions to receive equal average trust scores. If not, order and consistency of accurate suggestions would appear to matter. In some cases, we found that order does not matter. For example, there is no significant difference between groups '**110**', '**101**' and '**011**' ($p = 0.797$, $p = 0.155$, and $p = 0.114$ respectively). However, group **'111'** scored significantly higher than any of the groups with two accurate sessions, supporting H3.

In other cases, the importance of reliability and validity does influence trust averages, leading to significant differences between groups with equal correct suggestions. This is especially the case for groups that received accurate support from the intelligent system in only one of the three sessions. Shifting accurate system behavior by one session did not lead to a significant difference, i.e., neither group '**001**' and '**010**' ($p = 0.196$) nor group '**010**' and '**100**' ($p = 0.134$) differ from each other. However, group '**100**' was found to have a significantly higher average trust score than group '**001**' ($p = 0.013$). This suggests that a first good impression is significantly better for trust development than a repair through correct advice at a later point in time, supporting H4.

Additionally, average trust scores in group '**000**' did not differ from the groups '**001**' or '**010**' ($p = 0.960$ and $p = 0.269$ respectively), but were found to significantly differ from group '**100**' ($p = 0.027$). This partially supports H2. Additionally, this once again corroborates that the first good impression can make all the difference. Missing this opportunity for trust development in a first session causes later possible trust recovery to be futile. In fact, group '**100**' and '**011**' do not differ significantly in average trust scores ($p = 0.325$), even though the latter group corresponds to more correct suggestions than the former, underlining this finding further.

3.4.5 Trust Recovery is Possible

A further understanding of group differences can be derived from session differences within groups. The results of session comparisons per group can be found in Table 3.8.

Following H5, we expect trust to be significantly lower for an inaccurate session after it follows an accurate session. This is supported: we find this trust loss for groups '010', '100', '101', and '110'. The one exception to 'first impressions matter' and the one comparison that had an unexpected significant results, was within group '**011**'. While trust increase between session one and two was expected, trust increased further between session two and three, therefore, not supporting H6. A potential explanation can be that the impression of an improving system positively influenced perceived reliability of the system, leading to increased trust in the system. One possible explanation is that the impression of a learning system leads users to accept an initial fault when the system improves [237].

Table 3.8: Results of mixed ANOVA for average trust scores within groups between sessions. Green cells imply a significant difference between sessions. ↗ implies trust increased between the compared sessions, ↘ indicates trust decreased.

Group	S1-S2	S1-S3	S2-S3
000			
001		↗ **	↗ ***
010	↗ ***		↘ ***
011	↗ ***	↗ ***	↗ *
100	↘ **	↘ **	
101	↘ ***		↗ ***
110		↘ ***	↘ ***
111			

3.4.6 Dispositional Factors have Little Influence

Although system interactions influence trust development greatly, certain dispositional factors also shape trust evolution. These factors include for example age, gender, and country of origin [233]. To investigate **RQ4**, we gathered participant information for these factors, as well as the following: level of education, whether they studied computer science, their affinity with technology [176], and their propensity to trust [179].

We used linear mixed effects models to compare the influence of different dispositional factors. The fixed effects were set to "group * session", since the mixed ANOVA results from the analysis displayed in Table 3.7 and Table 3.8 showed a very strong interaction effect between experimental groups and sessions. Models with different added random effect variables were compared using ANOVA.

We found that out of all measured dispositional elements, two factors have a significant influence on trust evolution: age of the participant ($p = 0.006$) and their affinity with technology ($p = 0.012$). However, these traits only show a small effect ($sd = 0.24$ and $sd = 0.21$ respectively). Therefore, H7 was only partially supported.

The summarized results of our tested hypotheses can be found in Table 3.9.

3.4.7 Trust Evolves Slowly

The most detailed overview of trust scores can be found in Figure 3.3. In addition to the results of our hypothesis testing, we want to highlight interesting trends in the observed pace of trust formation. Many session comparisons with the same provided accuracy did not show a significant difference, but did show a trend in the expected direction. An example of this is group '**000**', where there is no significant difference, but trust drops

Table 3.9: Results of tested hypotheses.

Hypothesis	Result
H1. Increased amounts of accurate advice leads to more user reliance, while inaccurate advice will lower dependence on the intelligent system	Partially supported
H2. Consistent inaccurate advice leads to significantly lower trust than inaccurate consistent advice.	Partially supported
H3. Trust is significantly higher for users that receive consistent accurate advice.	Supported
H4. Inaccurate advice is more harmful in earlier sessions rather than later sessions.	Supported
H5. Trust is lost significantly if an inaccurate session follows an accurate session.	Supported
H6. Trust does not recover significantly when consistent accurate advice follows an inaccurate first impression.	Not supported
H7. The dispositional factors of gender, age, culture, experience with computer science, propensity to trust, and affinity with technology all influence trust formation across multiple sessions.	Partially supported

slightly over the sessions. Every group that has two inaccurate session suggestions shows a downward trend for the second incorrect session, no matter the order of accuracy. For positive trends, this is only the case when two accurate sessions are presented sequentially. Results from group '**101**' even show that trust between session one and three shows a downward trend. While these results are not significant differences, all found differences are in the expected direction.

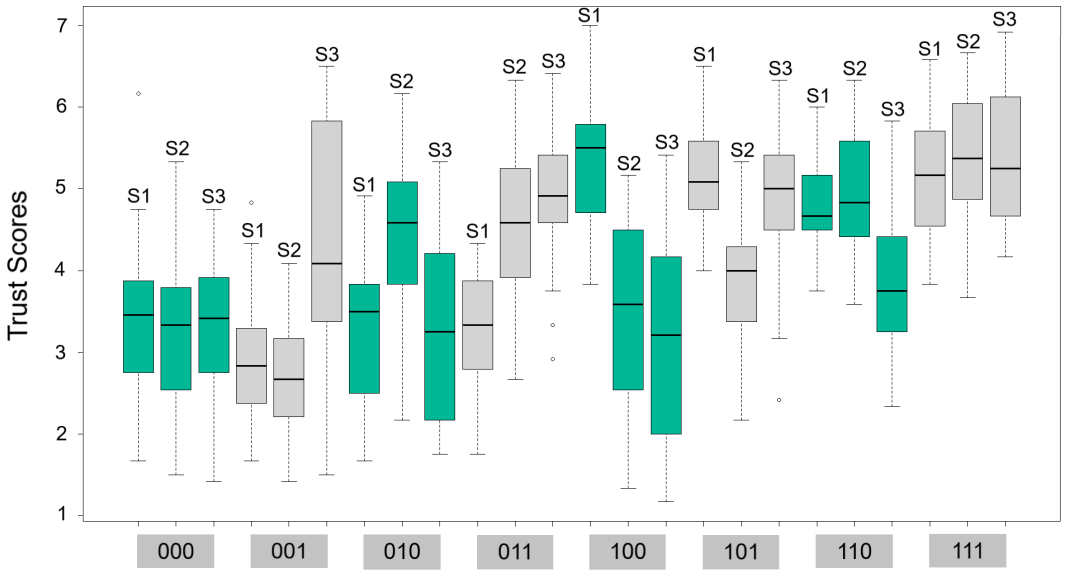


Fig. 3.3: Boxplots representing trust scores (*x-axis*) per session, across each experimental group (y-axis).

3.5 Discussion And Future Work

Our study revealed important interactions between trust formation and accuracy during intelligent system usage. Our work consolidates and complements previous studies of trust in HCI, and provides further insights on trust formation and evolution over time. In this section, we discuss our key findings and present further research directions that we believe are necessary to further understand user trust formation in intelligent systems.

3.5.1 Result Discussion

User Success: The fraction of correct houses found by the participants depended on two factors: system accuracy and difficulty of the scenarios. Particularly, difficult scenarios were more likely to be answered wrong, as were scenarios where the system gave the wrong advice. The fact that scenario complexity does influence user accuracy in session 1 and 3 seems to be caused by a learning effect: once users are used to the task in the first scenario, the difference between three and five constraints had less of an influence.

System reliance: Participants seemed to especially rely less on the system when they were in group **'000'**. Longer negative experience over time influenced their usage to decrease, especially in the last session. This indicated that even opening the system was not worth their effort. Submission reliance on the other hand had a closer correlation with system accuracy. Intuitively, groups with more accurate suggestions were more likely to submit those suggestions.

First impressions: The importance of first impressions in intelligent system interactions has been reflected by recent work [141, 366]. Our findings corroborate this understanding of Human-AI interaction. However, in contrast to related work that has primarily looked at trust formation within a single session, we measured trust development across multiple system interactions. We found that first impressions are not only important within a session, but also between sessions and over time. Interestingly, this is mostly the case when the system only had one accurate session. When the system provided two accurate sessions, there was no difference in trust values in all possible session orders (**110**, **101**, and **011**), even when the first session was inaccurate. More research is needed further investigate if increased system reliability, i.e., being accurate *most* of the time, indeed trumps the importance of first impressions in trust formation.

Trust recovery: We found that trust recovery is possible when a first system mistake is consistently corrected in later sessions. One possible explanation lies in the learning capabilities of the system, Earlier studies found that the impression of a learning system could lead to higher levels of trust [237], in line with our findings. As such, an interesting follow-up study can focus on features influencing perceived intelligence and how it influences trust formation.

Slow trust changes: The reported trust scores showed upward and downward trends when the system showed consistently accurate and inaccurate support respectively. These

slow but steady trends are reflected in research in the domain of autonomous cars [50]. We advanced the current understanding of trust evolution by introducing three consecutive sessions for each user. By expanding the number of sessions in future work, it can become apparent when and whether these trends become significant and/or plateau to a steady trust score in the longer-term.

Influence of dispositional factors: We found that age and affinity with technology influenced user's reported trust scores to an extent. While this is in line with earlier work [233, 437], other factors did not have a significant impact, including level of education, country of origin, gender, and propensity to trust. There are various possible explanations for these results. Firstly, as intelligent systems are becoming more pervasive, people from all levels of education come in touch with intelligent systems. The lack of significant effect of country of origin can be due to our sample: most of the participants were from Europe. It is possible that inter-continent comparison results in less effect than a comparison between continents. Finally, gender and propensity to trust did not have a significant effect. One possible explanation can be that participants did (not) experience the system to be intelligent. As the system starts to show more human-like traits, mental models related to trust in humans are more likely to be activated. We measured propensity to trust *in humans*, which does not correlate with the trust in our system if it is not perceived as intelligent enough. Future research could include different levels of anthropomorphism and system intelligence, to investigate its influence on trust.

It is striking that most dispositional traits had little to no effect on reported trust scores. Potentially, dispositional traits become less important as system experience increases. Alternatively, dispositional traits could influence trusting behaviors more than trusting beliefs. More research is needed on the effect of dispositional trust factors over time.

3.5.2 Caveats and Limitations

We make important contributions by advancing the current understanding of trust formation in Human-AI interaction. To position our findings within the scope of our study, we discuss the caveats and limitations of this work.

Firstly, we did not distinguish how wrong the intelligent system was. Incorrect advice consisted of a randomly assigned house that did not satisfy one or more requirements. The degree of incorrectness of an intelligent system can potentially influence trust formation. For example, a system that is very clearly wrong in its advice might lose user trust earlier than a system that is just slightly off. We aim to explore this in our imminent future work. The perceived utility of the system can also vary; adding more items in the search space could relate to more time saved by using system advice, while a larger bonus may also increase system usage.

Our focus in this work was on self-reported trusting beliefs. This is a direct measure, but can be subject to a reporting bias. Behavioral analysis, for example exploring whether

the participant heeded the advice, was used to corroborate our findings. However, further analysis can explore trusting behavior of users in comparison to trusting beliefs.

It is important to note that participants in our study were primarily European and fairly educated. The sample size of around 25 participants per experimental condition can limit the generalisability of our findings to other populations. Finally, as with much trust research, it can be questioned whether findings achieved in online studies can be replicated in real life. Experiments with intelligent systems being used in real life can both provide longer research windows to see if trends in trust formation over time become significant, as well as check the validity of online studies.

3.5.3 Implications and Future Work

Complex machine learning models and intelligent systems are currently being deployed in several critical domains, albeit as functional black-boxes. When human interaction with such systems, particularly in the first iteration, results in a sub-par experience, system adoption can be gravely affected. Impressions of a learning system can increase trust in the system, but only when the system actually appears to learn the correct behavior. Given that trust evolves slowly, system designers should focus on consistent behavior over time. Subsequently, system designers could benefit from trying other trust recovery mechanisms, especially when the user group is younger or has less affinity with technology.

Consistent system behavior over time can be investigated over a longer period of time with more sessions, to see if our results hold when users become used to the system and have calibrated their trust according to their experience.

In our work, we focused on the self-reported trust scores of users, or trusting beliefs. This in fact is only one aspect of trust: trust can for example be formalized as a disposition, attitude, belief, intention, or as behavior [335]. For example, while trusting belief usually has more emphasis on integrity of the trustee, trusting behavior focuses more on integrity and benevolence of the trustee [335]. There have been early results that suggest a mismatch between trust beliefs and trust behavior [489], which needs further investigation. We mainly focused on lack of accuracy as a cause for trust breakdown and improved accuracy as a form of trust recovery. To prevent trust loss in case of inaccurate AI support, different strategies for trust recovery can be deployed besides improving system accuracy. More research is needed into the effectiveness of such strategies related to different kinds of errors [130, 492].

3.6 Conclusion

Appropriate trust in intelligent systems is vital for successful and correct usage. Trust is not a static concept, but evolves during interactions over time. We presented a crowdsourcing study on the influence of system accuracy on trust formation over time. Answering **RQ1**, we find that accuracy explains opening of and using intelligent system advice.

Following **RQ2**, inconsistent accuracy of advice influences trust formation. Specifically, inaccurate advice leads to trust loss, earlier inaccuracy is more harmful than inaccurate advise in later sessions, and trust is significantly higher for users that receive consistent accurate advice. Session-wise comparison resulted in the answer for **RQ3**: inaccurate advice harms trust formation when it follows an accurate session and trust can be recovered after an initial inaccurate advice if the system provide accurate advice afterwards. With regards to influence of dispositional factors researched in **RQ4**, we discover that participant's age and affinity with technology have a small influence on trust formation. We identified the influence and importance of accuracy for trust formation and point to further research avenues on trust formation, trust break, and trust repair. In sum, this study provides first insights into trust development in response to system performance over multiple system interactions. Hence, it provides a first building block to understand this important and timely topic.

Taxonomy of Trust-Relevant Failures and Mitigation Strategies

Suzanne Tolmeijer[1], Astrid Weiss[2], Marc Hanheide[3], Felix Lindner[4], Thomas M. Powers[5], Clare Dixon[6], and Myrthe L. Tielman[7]

[1] University of Zurich, Switzerland
[2] Vienna University of Technology
[3] University of Lincoln
[4] Ulm University
[5] University of Delaware
[6] University of Liverpool
[7] Delft University of Technology

Abstract. We develop a taxonomy that categorizes HRI failure types and their impact on trust to structure the broad range of knowledge contributions. We further identify research gaps in order to support fellow researchers in the development of trustworthy robots. Studying trust repair in HRI has only recently been given more interest and we propose a taxonomy of potential trust violations and suitable repair strategies to support researchers during the development of interaction scenarios. The taxonomy distinguishes four failure types: Design, System, Expectation, and User failures and outlines potential mitigation strategies. Based on these failures, strategies for autonomous failure detection and repair are presented, employing explanation, verification and validation techniques. Finally, a research agenda for HRI is outlined, discussing identified gaps related to the relation of failures and HR-trust.

4.1 Introduction

Trust is an important component to ensure successful diffusion and uptake of human-robotic systems interaction in society. Trust in and trustworthiness of these systems have been considered important for long-term interaction, collaboration, and acceptance [300]. However, how should we design and implement trustworthy systems? Software engineering techniques such as verification and validation can be used to ensure that the system conforms to its requirements (verification) and the system meets the need of the stakeholder (validation). This improves reliability, safety and trustworthiness of the systems (see for example [184, 519]) and will help mitigate some of the failures leading to loss of trust.

Does the HRI community currently have sufficient knowledge of what makes a system trustworthy to be able to design robots as such? Human responses towards robotic systems are very complex in their nature and depend on many factors, such as the morphology and behavior of the system and the context in which they are deployed. Therefore, in order to design trustworthy robots, we have to base our design decision on detailed knowledge of (1) how humans react towards robots and (2) how robot features might foster or harm trust.

The challenge becomes more complex as trust has both static and dynamic components in human-robot interaction. Static components such as gender do not change, but dynamic components related to the system can be influenced [234]. We need to systematically structure the knowledge on trust that has been gained so far; it influences our design choices, also when an interaction is unsuccessful and possible negative effects need to be mitigated. The aspects of trust repair and trust violations have been understudied in the field of HRI [38]. Trust repair can be understood as the activity of rebuilding trust after one party breaks the trust of the other, i.e. after a trust violation. But what causes these trust violations and how can trust be repaired after they occur?

In this paper, we present a taxonomy of trust-relevant failures and mitigation strategies, based on literature as well as empirical data from known real-world use cases. Becoming aware of the fundamental need to structure our knowledge on how to build trustworthy systems, the discussion of this taxonomy started during a seminar where the authors met. The authors of this paper, who all have different disciplinary backgrounds ranging from philosophy and AI to mathematics and logic, analyzed the state of the art on trust research with respect to their disciplinary background. We propose a *taxonomy* that enables fellow researchers to incorporate mitigation strategies into their systems to recover from failure situations that potentially harm trust.

The inspiration for the taxonomy stems from so-called risk tables [64]. By definition, a risk equals uncertainty plus damage, in our case damage of trust [418]. In analyzing risks, one is attempting to envision how a scenario will play out if a certain course of action (or inaction) is undertaken. Therefore, a risk analysis always starts from three basic questions: (i) What can happen? (i.e., What can go wrong), (ii) How likely is it that that will happen?, and (iii) If it does happen, what are the consequences? [542] Classical risk tables visualize this information, e.g. the risk of getting a specific disease. We present an overview for failure situations in HRI that can harm trust in the robotic system, and offer robot designers mitigation strategies to (1) avoid or (2) recover from failure and reestablish trust.

We also outline explanation-based approaches, as well as validation and verification techniques that can be used to formalize our taxonomy in order to build trustworthy human-robot interactions.

4.2 Related work

Trust is a valued feature of individual human relationships which also enables social cohesion. Its dimensions have been studied by several disciplines, yielding results that both guide and limit the extent to which robot trust may be developed.

4.2.1 Approaches to Trust

The *psychology of trust* focuses on interpersonal relationships. The development of trust between persons typically follows familiarity, is concomitant with dependence, and in close personal relationships is associated with both behavioral predictability and the attribution of beneficent motives [412]. Risk regulation [353] allows the trusting agent to temper the degree of vulnerability to the party being trusted. Different kinds of trust attach to agents, depending on expectations and expertise. While the neurochemistry of trust is not well understood, it is assumed that trust can be understood both as a brain process and an emotional process [479].

The *ethics of trust* has been analyzed as necessary to economic exchange [26], friendships [482], and even the Hobbesian civil state itself [232]. Rusbult et al. [420] identified accommodation processes that allow close relationships to survive otherwise trust-breaking failures of expectations, e.g., via charitable interpretations of motives. Actions (commissions) that fail expectations and thus damage trust are, according to some [163], worse than failures to act (omissions). However, psychologists Tversky and Kahneman [497] as well as other ethicists find in this to be an omission bias, since the consequences of (not) acting can be the same. Hence from a consequentialist perspective, the damage to trust in the case of human failures ought to be similar. However, Malle et al. have shown that for robot failures, there is an asymmetry of blame–that humans blame robots more for failures of inaction than of action [323, 325].

Turning to *trust in robots*, we see the potential for overlap and contrast with the psychology, ethics, and pragmatics of trust between humans. Prior to the development of complex behaviors in robots, many philosophers would have insisted that trusting robots is more like trusting a tool than another person.With some conceptual flexibility, we can see that trusting robots has elements of both sorts. Studies on trust within robotics have mainly been motivated by the literature on trust in automation [235, 288, 294], which operates with a conceptualization of trust as mere reliance. According to this stance, trust is a domain-specific relation between the human and the robotic system involved. We follow this stance for our proposed trust taxonomy and define trust in robotic systems, in accordance to Lewis and colleagues [300], as a predictive belief or assumption about what will occur given the performance, process, or purpose of the robot. The definition of trust as appropriate reliance also stresses the importance of trust in situations involving risk and uncertainty. Humans who misplace trust, understood as both under- and over-reliance, might be exposed to serious danger, which is the reason safety concerns are of high consideration. In our understanding of trust as reliance, we consider the robotic system as tools intended for accomplishing certain ends. Other dimensions of trust, such as institutional trust are intentionally excluded, as our taxonomy should serve as robot-centered knowledge base.

4.2.2 Modeling Trust

The aim of defining/modeling trust in HRI is nothing new. Billings proposed a three-factor model of trust in robots, including *human* characteristics such as ability and personality, *environmental* characteristics such as task and team, and *robot* characteristics such as performance and attributes [61]. These three factors have also been identified in a meta-analysis on trust [211], where the authors stressed that too few studies have yet been conducted on environmental and human-related factors, although robot-related factors have been shown to affect trust the most.

Similarly, modeling trust from the perspective of risk has been considered before. Drawing on the model from organizational contexts by Mayer et al. [332] and the model on trust in automation by Lee and See [294], Wagner et al. [509] propose a trust model based on risk. They define trust as "a belief, held by the trustor, that the trustee will act in a manner that mitigates the trustor's risk in a situation in which the trustor has put its outcomes at risk" [509, p.26:4]. Trust is modeled in game-theoretic terms and, similar to what Hancock et al.[211] proposed, they highlight three important factors that influence trust-based decisions, namely the trustee, the trustor, and the situation. The model is also tested in an emergency experiment by Robinette et al. [417], where people tended to overtrust the robot despite half of them observing the same robot performing poorly in a navigation guidance task minutes before.

Based on the three-factor model by Hancock et al. [211], Hoff and Bashir [234] have also suggested a three-layered model in which these factors contribute to *dispositional*, *situational* and *learned* trust. They point out that age, gender, culture and personality are components of dispositional trust. Situational trust is shaped by internal and external variability, such as self-confidence and task difficulty. Learned trust consists of initial learned trust (e.g. expectations of the system) and dynamic learned trust. The latter is influenced by design features and system performance and influences the user's reliance on the system.

4.2.3 Trust, Failure, and Repair

The concepts of trust repair and trust violations have been understudied in the HRI literature so far. The need for research on trust in artificial agents in cases of inevitable failure has been highlighted as well [61]. Baker [38] surveys trust with a focus on trust violation and repair of human-robot interaction. For a successful recovery of trust, (perceived) shared intentions have shown to be important (cf. [138]). Even though from a scientific and engineering perspective we know that robots do not intend their behaviors in the same way as humans do, taking robots as intentional agents may aid users (psychologically) in attributing sufficient beneficence to their "motives" — at least insofar as this is necessary to engage with them. Following errors of automation, information related to limitations further aid in trust recovery. Hence, perceived benevolence may promote acceptance of a robot's changing behaviour [315], as with human interpersonal relationships [412].

In ongoing studies, several actions of trust repair have been proposed, including apologies, promises, internal or external attribution, and the showing of consistent series of trustworthy actions [38, 130]. In an emergency setting, where an apology right after violated trust has not recovered trust, an apology right before the next trust decision point has repaired trust. Promises lead to a better trust recovery than apologies, and in general, the message timing and exact content was shown to be crucial [416].

Thus, studies show that trust harmed by untrustworthy behaviour of a robot can be restored when people encounter a consistent series of trustworthy actions. However, trust harmed by deception and the same untrustworthy actions never fully recovers, even with actions of trust repair [444]. Additionally, a promise to change behavior can significantly speed the trust recovery process, but prior deception harms the effectiveness of a promise.

Studies on trust violation and repair take into account the evolving nature of trust, where trust is seen as something that changes over time. For example, Desai et al. [141] and Sebo et al. [447] researched robot failure and its influence on dynamic trust during one interaction. However, it has been outlined that long-term studies exploring the transient nature of trust are missing in the literature [300]. For example, how does trust change with increasing familiarity of the user with robots? Also due to their little employment in society, long-term studies have not been conducted so far.

Nordqvist and Lindblom [364] analyze trustworthiness of industrial robots with an operators' experience framework. The evaluation framework consists of the factors ability, benevolence, integrity, perceived safety, time on task and errors, where in total 12 user experience (UX) goals were characterized, 2 for each component. For each UX goal, data collection methods were selected and mixed, including observations, video recordings interviews, and Likert scales. Interestingly, major identified reasons for limited trust were communication problems during collaboration resulting in participant's uncertainty of their own ability to collaborate with the robot. The communication problems were strongly linked to the interface design. Further, the participants initially had confidence in the robot itself, but were insecure of their own ability to collaborate due to their inability to predict the robot's intentions and instructions.

In an online survey, Brooks et al. [75] explored people's reactions to failures in autonomous robots, namely a vacuum cleaner and a self-driving taxi, by manipulating four variables: context risk, failure severity, task support and human support. Participants' perceptions of an erroneous robot became less negative when it deployed a mitigation strategy, either by prompting task support, human support or both. However, the authors reported an interesting but non-significant tendency showing a preference for both task and human support in high severity situations, and a preference for only task support in low severity situations.

4.3 Proposal

We propose a taxonomy of failure types that can influence trust during Human-Robot Interaction. For each failure type, different mitigation strategies are suggested. While De Visser et al. [130] stress the importance of trust repair and list possible mitigation strategies, these strategies have not been linked to different failure types before. As mentioned by Baker et al. [38], models of human-automation and human-human trust are a helpful starting point, but do not account for the complexities of building and maintaining trust in HRI. A taxonomy for trust repair in HRI does not exist, but a framework for rebuilding trust in automation has been proposed by Marinaccio et al. [329]. It follows a similar intention: providing recommended trust repair strategies depending on the violation committed. However, they base their framework on the error taxonomy of Reason [410] which does not account for the interactive nature of HRI. Furthermore, human error taxonomies such as [451, 470] focus mostly on human error, while our taxonomy takes a holistic approach by including errors by other actors such as the system('s designer).

4.3.1 The Taxonomy

As a starting point for our discussions, we defined trust as "a person's willingness to rely on a robot to carry out its duties". As HRI involves two different actors, namely the robotic system and the human interacting with it, we base our taxonomy on a first fundamental distinction: who performed a type of action which caused a break of trust, (1) the system or (2) the user. Next, we distinguished the failure type (i.e. categorization of the actions into different types of failure). We differentiate four different failure types with respect to their impact on trust and the related mitigation strategies: (1) Design, (2) System, (3) Expectation, and (4) User (see Table 4.1 for condensed failure type descriptions).

Design. Imagine you have designed a robotic system in a specific way (in terms of behaviour, appearance, dialogue and so on) to the best of your knowledge. While in the real world the system behaves exactly the way you intended it to, it turns out that you made design choices that were not ideal for the HRI. For example, a specific function that you added to the robot is not used as often because the command is not as intuitive for the user as you thought, which influences the trust the user has in the system. A user misinterpreting the system's output because of its design; not understanding the interaction or not knowing about certain functionality when they should have are all considered Design failures. These failures are limited to the target audience of the system, as for Design failures the system's behaviour *should* be different in retrospect.

System. When a System failure occurs, the system does not act as intended. For example, the robot stops in the middle of a room during a navigation task without a reason, or stops a scanning task because its scanner malfunctions. In other words, the system does not do what it should, e.g. because of a system crash. The distinction can be made between a hardware and software failure.

Table 4.1: Types of actions which cause a loss of trust: we call these failures

Failure type:	**Action by**	**Meant to act this way**	**In retrospect, should actor behave this way?**	**Description**
Design	System	Yes	No	System does what it's been made to do, but in retrospect the system should not actually behave this way
System	System	No	No	System doesn't do what it's been made to do
Expectation	System	Yes	Yes	System does what it's been made to do, but user expected something different to happen. In retrospect system should still behave this way
User	User	No	If design fail: yes If expectation fail: no	User behaves in a way they are not supposed to. (Only a problem if leading to other type of failure)

Expectations. Trust in technological systems is typically concerned with the human's expectations of the system. With an Expectation failure, the system acts as intended, but defies the user's expectation. For example, when the user expects a robot to turn while observing a room, but the robot does not need to do so, the systems performs as it should but confuses the user. This is an example of an omission failure: the robot does not act when the user expects that it will. The opposite of this is a commission failure: the robot does something the user does not expect, e.g. start moving in the middle of an interaction because it needs to charge its battery. Expectation failures are different from Design failures in that for an Expectation failure the system should in retrospect still behave the same, while in case of a Design failure it should not. In case of the robot turning, the turning is an Expectation failure. However, there is probably a related Design failure as the robot does not explain its actions to the user properly. In this example, the Design failure is what leads to the Expectation failure.

User. In this last category the user interacts with the system in a way that he/she was not supposed to do, e.g. disturbing or sabotaging the robot (intentional) or standing in the robot's way so it cannot move (unintentional). This type of failure can be caused either by a Design failure or an Expectation failure which influences its impact on trust and potential mitigation strategies. While an Expectation failure deals with what the user

Table 4.2: Risk Analysis of Failure leading to loss of Trust (cf. Sec. 4.3.2)

Failure	**Probability**	**Impact on trust**	**Risk score**	**Mitigation strategy**
Design failure	3	2	6	ID, E, A
System failure				
Hardware	1*	3	3	E, A, F, Alt
Software	3*	3	9	E, A, F, Alt
Expectation failure				
Commission failure	2	4	8	E, A, ID, T
Omission failure	3	2	6	E, A, ID, T
User failure				
Intentional	2*	1	2	J, ID, Emo, Auth
Unintentional	2	3	6	T, ID

Probability scores: 1 = 1 occurrences in about 1000 interactions, 2 = 1 in 100, 3 = 1 in 10, 4 = likely in every interaction episode.
Impact scores: 1 = minor impact (negligible) to 4 = fatal impact (potential loss of trust and further use).
Mitigation strategies: ID = Interaction design; E = Explanation; A = Apology; F = Fix; J = Ask for justification; Emo = show emotion; Auth = Involve authority figure; Alt = Propose alternative; T = Training

expects the robot to do, a User failure is about what the users themselves do. Of course, Expectation failures could lead to unintentional User failures.

Combining all these failure types gives our foundation of the taxonomy shown in Table 4.2, including mitigation strategies that potentially repair the broken trust. This table is designed to resemble risk tables [64], also aiming to quantify the *Probability* of a failure occurring and the estimated *Impact* it will have on trust. In line with risk assessment practice, a *Risk score* is computed by multiplying the probability and impact scores, providing an indication of the priority for suitable mitigation strategies. These scores are system- and scenario-specific. To show how such a risk table can be used in a HRI context, the scores in Table 4.2 are from a real world use case.

4.3.2 Trust loss as a risk: A Case-Study

We present the following interactive system as a case-study in this paper, to show how our proposed taxonomy can be used for a real-world use case. In this example case [213, 218], an autonomous mobile robot has been deployed in a care home for a total of just over

a year, in the context of the STRANDS project[8]. This experiment was split over three individual deployments, following an iterative design principle, spread over a duration of three years. Here, the robot served as a mobile info-terminal and was also engaged in occupational therapy sessions. It was left without any technician or researcher on site, interaction with visitors and residents in the care home was without explicit solicitation by any experimenter.

Rich data sets, comprising task and error logs [218], user demographics [226], and navigation failures [135] have been obtained from these deployments, and analysed for the case study for this paper.

Tab. 4.2 presents the results of this case study analysis, in terms of *Probability* and *Impact* scores derived from the retrospective analysis of the data sets from the deployments. It shall be noted that this constitutes merely a case study, based on available data, allowing only some scores to be robustly computed from logs, while others have to be informed guesses, based on the authors' experience. For transparency, we have marked scores that are estimated from available data sets with an asterisk (*).

Probability and Detection In the specific instance, a variety of problems were detected automatically, such as navigation issues [135], forceful pushes to the robot, and hardware failures. Consequently, many failure types can be detected from system logs and from dedicated anomaly or failure detection modules that allow to estimate the probability of them occurring. In our case study of the STRANDS system, we analysed logs covering a cumulative deployment of over a year and employ some "Back-of-the-envelope" (BoE) calculations to derive the probability score. Given that the probability score is only intended to give an indication of the magnitude of a specific failure class, a BoE is most adequate for this assessment. The system data in [226] indicated that there were about 3.5 interactions per operational hour (i.e. time the robot is not resting or charging) with users that are actively using the robot. We shall take this estimate as the baseline for our BoE approximation.

An analysis of software failures, in particular navigation failures (which account for more than 99% of all software-related issues in this particular use case) in [135] reveals that in 1605 instances the robot had to ask for help as it could not recover from a navigation problem, making its failure obvious to the interacting humans, and hence potentially having an impact on trust. Thus, we observed such failure about every 2 hours of autonomous operation, leading to a ratio of 7 : 1 for Software failures to interactions, leading to a *Probability* score of "3" in Tab. 4.2. Most scores in Tab. 4.2 were calculated in a similar fashion: hardware failures were counted (e.g. snapper drive belt, failed encoder) as well as intentional User failures. In the case of the latter, by counting the number of forceful robot pushes and deliberate tampering, "intentional User failure" was observed in about 1 out of 200 interactions, scoring "2".

[8] http://strands-project.eu/

The other probability scores are much harder to obtain in a *post-mortem* analysis of long-term deployment, and require more focused studies, e.g. [219], involving the users directly. For instance, [219] revealed some of the Design failures that lead to the iterative improvements between annual deployments.

Impact To assess the impact of individual failures, we base our assessment of a qualitative analysis in the context of the care deployment within the STRANDS project [188, 219, 220]. The assessment is not an exact science; within this case study we do not aim for a comprehensive analysis of this STRANDS system, but rather present the concepts of considering trust loss as a risk open to a systematic analysis. For instance, feedback from on-site interviews showed that commission failures have a very high impact on trust. As an example, we quote a participant, who complained that the robot appears "stupid", because it would "start talking to a wall", a consequence of misclassification leading to a commission failure. However, establishing a robust scoring system for impact of trust that has wider applicability is one of the areas of future research.

4.3.3 Mitigation Strategies

Depending on the type of failure that has taken place, there are different possible mitigation strategies that can help regain the trust of the user. Given the interaction between different failure types, mitigation strategies for the initial failure type should be applied first. For example, if an Expectation failure was caused by a Design failure, the Design failure should be considered first. For Design, System and Expectation failures the following mitigation strategies can be used:

Fix. When a System failure occurs, be it hardware or software, the problem needs to be fixed. This is a very practical mitigation strategy to ensure the issue does not occur again and only applies to System failures.

Interaction Design. While it is intuitive that Interaction Design is important to foster trust, it can also be a tool in reestablishing trust. However, we can assume that once trust is broken due to a Design failure, the redesign of the system becomes even more challenging. As Lewicki and Wiethoff [299] explain restoring trust after a violation is a three-step process: (1) exchanging information about the perceived trust violation, (2) willingness to forgive the violator, and (3) reaffirm their commitment. Implicitly communicating all of these aspects to the same user with a change in interaction design will be hardly possible. However, improving trust through the interaction design for other prospective users will still be a viable way to go. Proper design allows for smooth interactions and substantial research is available in HRI on understanding robot-related factors affecting trust in the interaction design, such as social skills [222], robot role [198], and communication style [408]. Hancock et al. [211] provide a detailed overview on HRI studies on the impact of robot design features on trust in HRI.

Explanations. Explanations for the end user can be a suitable mean to repair trust. Methods, such as plan-based explanations related to previous decisions can be used. However, the correct level of detail of abstractions and human-comprehensible explanations are challenging. Explanations to end users do not necessarily need to be in natural language, but can use cues such as closed eyes, blinking lights, nodding head etc. Overall, the aim of explanations should be to increase transparency and understandability in order to repair trust in a failure situation.

Apology. Once a trust failure occurs, it is essential to recognize that trust has been broken and acknowledge that the failure that occurred was unpleasant for the user. Apologies are effective for trust violations related to the violator's competences (e.g. an error in planning or judgement) [281]. In human-human interaction, they are more effective than shifting the blame elsewhere. Once the human understands the effect was not intended and is not intended to happen in the future, the trust repair can start. Lee et al. showed that the apology strategy was most effective to mitigate perceptions of competence, closeness and likeability of a service robot [296].

Propose Alternative. In case of a system breakdown, the trust lost in the system can be minimized when alternatives are available. If possible, the system can propose a workaround the user can employ to still get the intended task done despite a System failure.

Our discussions on User Failures revealed that there is little to no research on how to mitigate this type of failure. We consider the following strategies as promising:

Ask the Human for Justification. When a user misbehaves, the response the system gives will influence future behavior of the user towards the system. If the user was not aware of any misbehavior, asking the user for justification of their actions can create awareness of their mistakes. We assume that unintentional negative behavior will not be repeated once the user becomes aware of it. Intentional misbehavior is harder to address, since the user acted purposefully. Asking a justification is intended to help the user realize the negative consequences of their actions.

Show Emotion. It is in our nature to anthropomorphize robots, for example by projecting a personality onto the robot or reading emotions into its output. When a user misbehaves, emotion can be a powerful tool to persuade the user to behave better. However, the impact of negative emotions displayed by a robot is understudied [249]. The only study we are aware of, in which a robot shows negative emotions - namely an aggressive movement pattern - could show that this was enough to reduce robot abuse [438].

Involve Authority Figure. Using authority is a persuasion mechanism [104] that can be useful to make sure users behave properly towards the robot. An example can be to alert

the owner of the robot or authorities. Research on children's abusive behaviour towards robots in shopping malls revealed, that children typically did not stop such misbehavior until their parents (their authority figure) stopped them or they got bored [76].

Training. For unintentional User failures, training can be a potential mitigation strategy to avoid repeated future failure situations. So far, little research has been done on how users can be trained in HRI (since research mostly focuses on how users can train robots [17]), but existing work shows that "training is essential" [81].

4.4 Autonomous Trust Repair

What we want to achieve in HRI at some point is autonomous trust repair, which implies both failure detection and failure mitigation is managed without human assistance. The first step towards this goal is failure detection: is something wrong with the system? Related to this is failure classification: once it is established something is wrong, the system needs to assess what is wrong. Finally, using this classification and the detected deviation from the plan the system had, an explanation can be presented in an attempt to repair the lost trust. In our opinion, this is a fundamental prerequisite: a robot needs to detect that a failure happened and an explanation to the end user should be the starting point for any mitigation strategy.

4.4.1 Failure Detection

Robots that interact with humans in the wild will at some point face failure situations, which can either be inflicted by the robot, the human, or by unexpected environmental events. However, dealing properly with failure situations from a robot-centered perspective is a challenging endeavour. Firstly, the robot has to detect that an error situation has occurred; secondly, it needs to analyze what kind of error situation occurred; thirdly, it needs to apply an error recovery strategy to get back into a safe interaction state.

What can be detected? Looking at our taxonomy in Table 4.1, the question arises which of those failure types can be detected by a robot itself (self-awareness) without further involvement of the user. The common definition of failure usually requires the exact knowledge and definition of a *failure case*, i.e., a formal definition of what constitutes a failure. In other words, the failure detection problem is considered a classification problem, where a model of the failure itself can either be defined or learned.

One way to do this is by using verification and validation techniques. Formal verification is a mathematical analysis of all behaviours of the robot or system using logics, and tools such as theorem provers or model checkers (see for example [105, 172]). Using model checking, a desirable property encoded in some logic is checked over a model, often a finite state transition system, to ensure that it holds on all paths through the system from an

initial state. Theorem proving involves a mathematical proof to show that the property expressed in some logic is a logical consequence of the system also expressed in logic. Simulation based testing utilises simulations of the robots and the environment, possibly including hardware in the loop, to facilitate large numbers of tests that may not be possible in the real world. Tools are used to automate the testing and analyse the coverage of the tests. End user experiments can be used to test aspects such as trustworthiness. Formal verification, simulation based testing and end user experiments can help improve the safety, reliability and trust in robotic systems [184, 519], as well as help mitigate as system failure (all), design failure and expectation failure (end user experiments).

However, this approach limits the ability to detect failures to properties that have been identified in the specification. A complementary approach relates to *anomaly detection* (e.g. recently surveyed in [202]). It aims to detect any deviation from a normal behaviour of a system, without necessarily classifying a problem. The identification of a potentially known problem can then be deferred to approaches to generation explanations, utilising domain knowledge as formally defined in the following section and also explored in [212].

4.4.2 Offering Explanations

Once a failure is detected and possibly classified, we consider explanations as one possibility for failure mitigation (see Tab. 4.2). Therefore, it is desirable to investigate how a robot can automatically generate explanations based on its perception and deliberation modules. According to Miller [344], explanations should be *contrastive*, *selective*, and *social*. Contrastive explanations (implicitly or explicitly) refer to situations different to the one to be explained. For instance, *Why does the robot do X?* should be understood as *Why does the robot do X rather than Y?* One way to generate contrastive explanations is by counterfactual analysis: the occurrence of some phenomenon X in situation S can be explained by a sufficiently altered situation S' where X does not occur (but Y does). Counterfactual explanations have recently been applied to generating explanations for plan failures [189], for explaining why an action plan contains a specific action [175], and to explain why an action plan does (not) adhere to moral principles [308]. These approaches only partially fulfill Miller's criteria of selectivity, though: although minimality criteria are considered, there are generally many possible explanations and it is not necessarily clear how to pick the most appropriate ones. Wang and colleagues [516] circumvent this challenge by generating explanations from Partially Observable Markov Decision Processes using a template-based approach. The downside of this approach is its being less generic and its requiring hand-crafted template modeling. Finally, Miller requires explanations to be social, that is, explanations should take the user's mental state (beliefs, desires etc.) into account. This requirement is a big challenge to the current state of the art of explanation generation.

4.4.3 Formalism for representing plans

A procedure for explaining failures can be based on the STRIPS formalism for planning [169]. STRIPS and its derivatives are widely used to describe a robot's deliberate actions and external events. A STRIPS model is a tuple $\langle P, s_0, s_g, O, pre, del, add\rangle$ with a set of propositions P, an initial state $s_0 \subseteq P$, a partial state $s_g \subseteq P$ called goal description, a set of operators O (actions and events), a function $pre\colon O \mapsto 2^P$ mapping each operator to a set of preconditions that must hold for the operator to be executable, a function $del\colon O \mapsto 2^P$ mapping each operator to a set of propositions to be deleted from the current world state as an effect of the operator's execution, and a function $add\colon O \mapsto 2^P$ mapping each operator to a set of propositions to be added to the current world state. The execution of operators thus triggers transitions from current world states to successor world states, where world states are sets of propositions. An operator o is *applicable* in a state s iff $pre(o) \subseteq s$. The successor state $s' = (s \setminus del(o)) \cup add(o)$ results from applying o in s. A state s is a goal state if $s_g \subseteq s$. We assume the existence of the empty action $\epsilon \in O$, which has an empty precondition, an empty delete list, and an empty add list.

As an example, consider a robot currently situated in the kitchen. It wants to move to the dining room. The applicable action operator $move(kitchen, diningroom)$ has precondition $\{in(kitchen)\}$. The action's effect is given by the delete list $\{in(kitchen)\}$ and the add list $\{in(diningroom)\}$. Hence, by performing the action $move(kitchen, diningroom)$ in state $s_0 = \{in(kitchen)\}$, the world state transitions from state s_0 to state $s_1 = (s_0 \setminus \{in(kitchen)\}) \cup \{in(diningroom)\} = \{in(diningroom)\}$.

4.4.4 Explaining failures from plans

Let $\pi = s_0 \rightarrow_{o_0} s_1 \rightarrow_{o_1} \ldots \rightarrow_{o_{n-1}} s_n$ be a course of actions and events o_i—also called a *plan*—originating from the initial state s_0 and terminating in some state s_n, which may (or may not) qualify as a *failure state* in the sense of the conceptualization outlined in Subsect. 4.3.1 and Tab. 4.2. In case of failure, we want to understand why the failure occurs by answering Why-questions about facts and actions:

1. Why does fact p (not) hold at time point t?
2. Why does the robot (not) perform action a at time point t?

As an example, consider the following case which involves an *expectation failure* of type *commission* and requires generating an answer to a question of type (2): *After the robot receives a navigation goal from the user, the robot suddenly starts turning to get a precise estimate of its current location via its front-mounted laser rangefinder.* The user expects the robot to immediately start moving towards the specified destination and thus wants to understand *Why does the robot start turning?* To see how an answer can be generated, first consider the robot's plan $\pi = s_0 \rightarrow_{tl} s_1 \rightarrow_{nd} s_2$, i.e., the robot plans to first make a turn to improve its localization (action tl) and then to navigate to the destination (action nd). Initially, the robot's pose estimate is poor (fact pe) and the robot

is not at the destination, i.e., $s_0 = \{pe\}$. The goal $s_g = \{d\}$ is to be at the destination. The precondition of the navigation action nd is that the robot has a good pose estimate (fact ge). Performing nd adds d to the state. The turn action tl has delete list $del(tl) = \{pe\}$ and add list $add(tl) = \{ge\}$. To explain why the robot is turning, counterfactual analysis is used: an inclusion-wise minimal subset $x \subseteq add(tl)$ of the add list of action tl is identified, such that if the facts in x were removed from $add(tl)$, then the final state of plan π would be no goal state. This is to say that x is a necessary means to the goal d. Clearly, removing fact ge from $add(tl)$ would make action nd inapplicable and thus fact d would be missing from the final state. Accordingly, the robot can explain *Turning around results in knowing where I am, and this is necessary for finally reaching the destination.*

4.4.5 Logics for Trust Loss Detection

One way to recognize whether trust was lost because of a failure, is by using logics to model and reason about trust loss. Logics for trust have been developed. In [227] the authors formalise the work of [86, 160]. In [86, 160], i (truster) trusts j (trustee) to do α (an action) with respect to φ (a goal) if and only if (1) i has the goal φ; (2) i believes that (a) j is capable to do α; (b) j, by doing α, will ensure φ; and (c) j intends to do α.

In [227] the notion of trust is reduced to more primitive concepts of belief, goal, capability and opportunity which is formalised in a logic of time, action, beliefs and chosen goal. Two kinds of trust are considered. Firstly, the truster believes that the trustee is going to act here and now (termed occurrent trust). Secondly, the truster believes that the trustee is going to act whenever some conditions are satisfied (dispositional trust). Only occurrent trust and qualitative aspects of trust are considered in [86, 160]. Two dynamic logic operators $After_{i:\alpha}$ and $Does_{i:\alpha}$ are proposed. The former gives the result of agent i's performing action α (its capabilites) and the latter about what holds after agent i does action α (what an agent does and intends to do). The modal operators Bel_i (agent i believes) and $Choice_i$ (agent i has chosen the goal) and the temporal operators G (always in the future) and F (at some future moment) are also used. Occurrent trust $OccTrust(i, j, \alpha, \varphi)$ is defined as follows:
$OccTrust(i, j, \alpha, \varphi) = Choice_i F\varphi \wedge Bel_i(Does_{j:\alpha}\top \wedge After_{j:\alpha}\varphi)$.
That is i trusts j to do α with respect to φ if and only if, i wants φ to be true at some point in the future and believes that j will ensure φ by doing action α. The authors argue that this may be too strong as j is going to do α immediately. This leads to the definition and formalisation of dispositional trust which is weaker than this. A complete axiomatisation is provided but complexity and decidability are not considered.

In [254, 255] the authors consider automated quantitative reasoning about trust via stochastic multi-agent systems. They formulate probabilistic rational temporal logic (PRTL*) as a combination of the probabilistic computation tree logic (PCTL*) with cognitive attitude operators (belief, goal, intention) and trust operators (competence, disposition and dependence). The resulting logic is, in general, undecidable but decidable

fragments are identified. The work has again been inspired by [160] and, as with our work, the focus is on trust between humans and robots/autonomous systems.

These logics could be used to model robotic trust scenarios to identify when and how the system is not trusted or trust is lost. The belief aspects from [86, 160] and modelled in the logics mentioned above seem to match the expectation failure type discussed above. However, they do not match the more complex models of trust as introduced in Section 4.2.2.

4.5 Future Work

Reflecting on existing HRI research on trust repair and the introduced taxonomy, as well as autonomous failure handling through explanation generation, verification and validation techniques lead us to identify research gaps we consider crucial to be further explored for successful trust failure classification and mitigation.

Mitigation of User Failures Our discussions identified the category of intentional and unintentional *User Failures* as up-to-now understudied with respect to mitigation strategies [245]. Mainly how robots could react if people intentionally cause errors, e.g. by covering sensors, giving wrong information or other ways of intentionally bullying the robot. We gave potential examples of mitigation strategies, namely calling an authority, showing emotions, and ask the person for justification. However, effects of robots showing negative emotions are in general understudied [249], and no systematic studies of the other strategies exist so far.

Impact of Failure Repetition Similarly, the impact of failure repetition is understudied, above all with respect to how it affects trust. Some studies on people's willingness to help robots after repeated failure indicate that repeatedly helping robots in need when the suggested repair strategy is successful does not reduce likability [37]. However, this does not give insights into how much overall trust is harmed. It will need long-term studies outside of laboratory experiments to get an ecologically valid grasp on how failure repetition affects trust. Subsequently, long-term in-the-wild studies, lasting several weeks to out-rule novelty effects [190], will be needed to assess the impact of familiarity with the robot. Studies on gracefully failing robots will substantially inform trustworthy HRI design.

Severity Rankings Failure classifications often come with severity rankings, such as the failure classification by Carlson and Murphy [85]. They classified physical failures according to severity (terminal failure: terminates the system's current mission; non-terminal failures: degrades its ability to perform its mission) and repairability (field repairable: repairable with tools that accompany the system in the field; nonfield repairable: cannot be repaired with tools that accompany the system in the field). For our approach we would

like to extend our taxonomy with a severity ranking with respect to the loss of trust. Similarly, to the impact of repetition, data from long-term field trials will be needed in order to add empirical evidence to our taxonomy.

Automated recognition of Trust Loss As mentioned before, the current logics that allow trust (loss) modeling are fairly simplistic. Furthermore, different cues in user behavior need to be distinguishable to detect trust loss of the user in the system. While automated detection of a failure is the first necessary step in failure mitigation, the next goal should be automated trust loss detection to be able to respond appropriately. As the STRANDS use case has shown, proper recognition and standardized scoring of trust loss could greatly benefit trust research in HRI.

4.6 Conclusion

In this paper, we aimed at consolidating the knowledge we have on trust and trust repair in HRI in a taxonomy with the aim to help fellow researchers developing trustworthy robots according to the state of the art. We aimed at specifically structuring potential failure situations from the robot as well as from the user perspective. Our efforts revealed that empirical research in HRI tries to more and more identify suitable mitigation strategies, but hardly considers the type of failure that caused the trust violation. We argue that a framing of failure situations will have an impact on trust repair and needs to be considered in future studies, but above all in future interaction designs. Moreover, we tried to outline how failure detection could be improved for future HRI, as well as the logics of verification of failure states. Future work in these areas will be essential to actual enable autonomous trust repair in HRI including autonomously generated suitable explanation strategies.

AI for Ethical Decision Making

This chapter is based on:

Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. ***Capable but Amoral? Comparing AI and Human Team Members in Ethical Decision Making****. Conditionally accepted (revise and resubmit) at CHI 2022: Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA*

Capable but Amoral? Comparing AI and Human Team Members in Ethical Decision Making

Suzanne Tolmeijer[1], Markus Christen[2], Serhiy Kandul[2], Markus Kneer[2], and Abraham Bernstein[1]

[1] Department of Informatics, University of Zurich, Switzerland
[2] Digital Society Initiative, University of Zurich, Switzerland

Abstract. While artificial intelligence (AI) is increasingly applied for decision-making processes, ethical decisions pose challenges for AI applications. Given that humans cannot always agree on the right thing to do, how would ethical decision-making by AI systems be perceived and how would responsibility be ascribed in human-AI teams? In this study, we investigate how the expert type (human vs. AI) and level of expert autonomy influence trust, perceived responsibility, and reliance. We find that participants consider humans to be more morally trustworthy but less capable than their AI equivalent. This shows in participants' reliance on AI: AI recommendations and decisions are accepted more often than the human expert's. However, AI team experts are perceived to be less responsible than humans, while programmers and sellers of AI systems are deemed partially responsible instead.

5.1 Introduction

The capabilities of artificial intelligence (AI) technology continue to grow. Increasingly, AI is being applied to support and even take over tasks from humans, ranging from creating new recipes [388] and co-creation of art [302] to HR decisions [377] and clinical decision making [295, 538]. This provides many possible benefits: tasks that are risky or challenging for humans, tasks that are done more efficiently by AI, or tasks that require specific AI skills such as pattern analysis in large data sets, could all be outsourced to AI. However, for implementations to become successful, users need to trust the system enough to be willing to use it. Depending on the domain and application, mixed results have been found on user trust in AI. One stream of research found signs of algorithmic appreciation: people believe AI performs at least as good, if not better, than human experts [16]. Especially lay people seem to trust an AI more in various cases, such as forecasts of song popularity or romantic attraction [310]. However, another set of experiments has shown indications of users experiencing algorithmic aversion. For instance, people lose trust in AI faster when it makes mistakes than when a human expert does [142]. Users are more likely to experience algorithmic aversion if they have incorrect expectations, experience a lack of decision control, and when AI suggestions go against the user's intuition [80]. All of the mentioned factors that can trigger algorithmic aversion depend on the decision domain and task type the AI performing in.

Not all tasks are made equally: while some are generally accepted to be outsourced to AI, others might pose ethical conundrums. As is often said in computer science, "garbage

in, garbage out": only when you give clear, unbiased, well-balanced data as input, can AI produce useful output. A ground-truth is needed, a guideline on what is correct or incorrect output. In the case of ethical decision making, this ground-truth poses a problem. After all, philosophers have discussed for centuries whether ethical behavior has more to do with intentions (such as Kant's deontological ethics [278]), outcomes (such as Bentham's consequentialism [57]), virtues (dating back to Aristotle and Confusius), or yet another approach altogether. A practical example of this clash of ethical preferences can be found in a program called COMPAS, which was created to predict recidivism in American inmates. Initial research found a strong bias against African American defendants [280]. However, upon further analysis, it turned out that it would have been mathematically impossible to adhere to people's different notions of what a more fair outcome would have been [311]. While research on implementing ethics in AI has being ongoing, but in a scattered and relatively limited matter [491].

Part of what makes ethical AI so difficult to implement in practice, is the important challenge of responsibility ascription — especially when a decision could lead to negative outcomes. Traditional concepts of ascribing responsibility do not apply to AI, since AI is not considered to be a moral agent, leading to the risk of a 'responsibility gap' [331]. In the context of ethical decision making for AI in severe contexts, such as with autonomous weapons systems, this had lead to the discussion of 'meaningful human control': AI should respond to input from human experts and every AI decision should be traceable to a human [430]. The importance of the human element to ensure legal compliance and ethical acceptability when using AI in security contexts is considered indispensable by stakeholders such as the ICRC [368]. In other words, people prefer to have a human responsible for the outcomes, so someone can be held accountable for mistakes emerging from AI decisions. Whether or not people perceive different parties involved in the AI system to be responsible is an ongoing topic of research. Generally, users assign more responsibility to parties that have more autonomy in decision making [244]. Different types of agency lead to different responsibility ascriptions, such as to the AI artifact, the designer, and the user of the system [271]. The assigned responsibility also depends on the role and autonomy the AI has.

AI can be applied in a *human-in-the-loop* (HITL) setting or a *human-on-the-loop* (HOTL) setting [355]. The former implies that the human has the main decision power but is assisted by the AI, while the latter means that the AI makes decisions but a human overseer can veto AI decisions and correct mistakes when they happen. We expect that the level of autonomy influences trust in the system as well as the responsibility assigned to the AI.

Eventually, perceptions of trust and responsibility lead to (the lack of) reliance on AI systems. Reliance implies that users are willing to follow the AI's decision or recommendation. Since trust guides reliance, AI systems should set correct expectations, leading to appropriate reliance [293]. Chiang and Yin [102] found that increasing people's

understanding of how machine learning performance depends on the task, led to less over-reliance. Responsibility also shapes reliance as long as it is unclear who is responsible and liable, users will be more hesitant to rely on AI [4].

No matter how theoretically sound a particular AI implementation is in respect to a particular ethical view, people's perceptions ultimately shape the reliance on and the success of the technology in practice. Therefore, empirical evaluation of the perception of AI in different domains gains importance. While there have been separate studies on trust in AI, responsibility ascription, and reliance on AI, to our knowledge, this combination of factors and their interaction have not been researched in an empirical setting for AI making ethical decisions. Especially in the context of human-AI teams, this combination of factors is vital to make the AI application a success in practice.

This work focuses on the perception of ethical decision making of AI for different levels of autonomy for scenarios in the search and rescue (SAR) and defense domain. Specifically, it focuses on trust placed in the AI and who is deemed responsible when humans and AI work as a team for ethical decision making. To this end, was had participants make ethical decision using a 2x2 experimental design, to research people's perception and reliance behavior for different factors: type of expert (human vs. AI) and level of autonomy (human-in-the-loop vs. human-on-the-loop). We have chosen two different ethical decision domains, because research has shown that different task domains trigger different ethical behavior associated with main ethical theories (such as deontological ethics or consequentialism) [103]. Thus, the task framing serves as control condition to ensure that not one single ethical theory dominates the decisions made. We present two different types of scenarios: the task either involves minimizing casualties (defence domain) vs. maximizing lives saved (search and rescue domain) and advice is pretested to not perceived to be clearly wrong. Since the Trolley Problem, the standard type of dilemma used for ethical decision making in severe contexts, is a simplistic sacrificial dilemma that lacks realism from a moral psychology perspective [48], we choose a more realistic approach: we include uncertainty regarding decision outcomes as a part of the dilemmas participants face in the experiment. We looked at how the mentioned factors influenced 1) trust placed in the human and AI expert, 2) perceived distribution of responsibility in the different settings, and 3) reliance on the expert's suggestion. This allowed us to investigate the following research questions:

- RQ1: How does reported trust in a human and AI expert compare for ethical decision making support?
- RQ2: How is responsibility attributed when interacting with a human or AI expert with different levels of autonomy (HITL vs. HOTL)?
- RQ3: How does reliance on human vs. AI advice compare?

Our results indicate that people perceive AI to be more capable than humans for the given tasks, but place higher moral trust in humans. The capable trust in AI is apparent

in participant reliance behavior: as they do more missions, they are more likely to take an AI's advice or accept an AI's decision than a human expert's. Additionally, an AI is considered to have less responsibility than human experts, while programmers and sellers of AI technology carry part of the responsibility instead. Our findings contribute to the research on human-AI teams and AI for ethical decision making, by presenting design implications of our findings.

5.2 Related Work

AI is different from other technology users have interacted with thus far, leading to new challenges in the design of human-AI interaction. Major challenges in designing AI are related to uncertainty about AI's capabilities and the output complexity that AI offers [537]. Perception of AI also differs from earlier technology because AI is still a fairly new technology for users to interact with and they are uncertain what it could do for them [456]. In this section, we summarize ongoing work, focusing on the difference in perception between humans and AI doing tasks, trust and perceived responsibility in AI, and how ethical decision making has been considered for AI thus far.

5.2.1 Perception of Human vs. AI Experts

AI is able to operate in a more autonomous fashion as its capabilities increase. Initial AI applications focused on decision support — AI can already support the clinical decision making process [538], group decision making [279], and advise on what to eat [480] or watch [350]. Now, applications are moving towards autonomous analysis of tasks, such as diagnosis based on medical images [204] or autonomous task execution like driving a car [260]. In the next ten years, AI is already expected to outperform humans in jobs such as translating languages, writing high-school essays, and driving a truck [195].

In this ongoing shift of tasks towards AI, comparing performance and perception of human experts with their potential AI counterpart is a logical next step. Depending on the specific algorithm, domain, and application case, different results have emerged from this comparison. While capabilities are slowly increasing, positive perception is not rising in the same manner. Especially when AI is applied in a ethical context, AI has the additional challenges of meeting social expectations on top of functional ones, leading to varied results in perception. General perception of AI has shown an increase in fears of loss of control and ethical concerns [162]. Specifically, people worry about the usefulness and fairness autonomous AI on a societal level, even though AI is considered at least equally capable as human [16]. On an individual user level, Chen et al. [96] found that while patients appreciated a human doctor remembering specifics of their case, they found it intrusive when an AI doctor did the same. Human experts are considered more fair than AI for *the same* recruitment decisions [360]. Human artwork is evaluated more highly

than AI artwork [405]. On the other hand, news articles written by AI and human news editors were considered equally credible [526].

One important factor that relates to perception of AI is trust. Following Lee and See [293, p 51], we define trust as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*". Trust in the system influences whether and how AI is used in practice. Framing of the system and capabilities of AI before usage highly impact acceptance and accuracy perceptions of users [285]. When trust in AI systems is higher than in human experts, this can lead to what Logg et al. [310] have dubbed *algorithmic appreciation*. In their study, they found that people use AI advice more than human advice, even when the system's process is opaque. Additionally, Thurman et al. [485] found that this effect also holds when the advice comes from human experts rather than just laypeople. While people sometimes worry about the consequences of autonomous AI, they still consider to be AI to be as good as or better than human experts [16]. One possible explanation is the machine heuristic, in which humans consider AI to be more objective and less ideology-biased than humans [474]. However, whether this also applies in ethical decision making has not been researched yet.

On the other hand, when people do not trust AI and prefer human experts, the literature speaks of *algorithmic aversion*. For example people are more sensitive to AI making mistakes than humans; it causes them to loose trust faster [142]. One way to overcome this aversion, is by framing the system to be a learning system [58]. In a literature review, Jussupow et al. [274] found that preference for human vs. AI depended on the expertise and social distance to the human expert, and agency, performance, capabilities, and human involvement in the training for AI expert systems. People had less algorithmic aversion for machines that performed more objective quantifiable tasks, but more when the task was considered more subjective [87]. Since ethical decision making could be considered more subjective, we consider the following hypothesis:

H1: People show more algorithmic aversion for AI making ethical decisions, implying they show less trust in AI compared to a human expert.

5.2.2 Perceived Responsibility of Autonomous AI

Part of the challenge of using autonomous AI is the ascription of responsibility of decision making. In terms of positive consequences of AI, responsibility can be hard to assign. An example of positive outcome responsibility is income resulting from the generation of art by AI systems. Epstein et al. [156] found that allocation of responsibility is influenced by perceptions of anthropomorphism of the system, which is partially influenced by the language used to describe the systems.

Responsibility of negative results is perceived differently. Research so far has shown that people are willing to assign moral blame to AI, especially when AI systems become more sophisticated [284]. However, compared to humans, the type of responsibility that

is assigned differs. AI receives similar blame and causal responsibility, but less moral responsibility: in bail decision making, human agents are ascribed higher levels of present-looking and forward-looking notions of responsibility [305]. In some cases, such as by younger adults, blames falls more on the programmer making the AI rather than the AI itself [180]. However, an individual programmer is not the only person influencing actions of the AI: "*Responsibility would need to be assigned collectively to all actors contributing to this AI system. But collective responsibility is a notoriously difficult concept, as being morally responsible requires moral agency, and it is not completely clear under which circumstances, if any, a collective qualifies as a moral agent*" [221, p 14]. This effect of responsibility diffusion has been researched in social psychology (e.g., [174, 354, 510]), but not yet in the contest of human-AI teams.

In addition to issues with perception of responsibility, the legal system is not equipped to deal with criminal liability of AI systems yet [369]. For instance, liability and data usage of AI creating news articles are currently becoming an issue [347]. Creating new legislation on AI responsibility that is considered fair, can benefit from a deeper understanding of responsibility assignment of lay people — something that is investigated in this study. We hypothesize the following:

H2 AI is perceived to be less responsible than a human expert. Level of autonomy has a larger influence on responsibility ascription for human experts than for AI.

5.2.3 Human-AI Team Configuration and Reliance

Humans and AI reason differently, leading to both parties having different strengths and weaknesses. Rather than aiming for AI to take over tasks completely, human-AI teaming could be a fruitful alternative to combine strengths and produce new possibilities for the future of work [267]. Especially in the context of meaningful human control, AI's cannot act independently for ethical decision making, but is preferred to be part of a team that includes humans as well. Human-AI teaming, also coined Human-Autonomy teaming, is recently gaining traction as a research field (see, e.g, [371]). In the context of the role of humans in the Human-AI collaboration, two prominent configurations of Human-AI teams have been discussed: human-in-the loop and human-on-the loop settings (see, e.g, [171, 247, 355]). The human-in-the loop configurations are characterized by an active involvement of a human at various stages of the process (higher degree of human control, less autonomy of an AI). The human-on-the loop configurations in contrast are characterized by rather passive involvement of a human in the process (lower degree of human control, higher autonomy of AI). In case of human-in-the-loop, AI functions as decision support system giving recommendations to a human in the team who then decides upon receiving such a recommendation. In case of human-on-the-loop, most of the decisions are delegated to AI and humans only monitor the AI and intervene if they deem it necessary.

The varying degree of human involvement in the human-AI teams might impact people's perception of AI, and therefore affect the degree of people's reliance on AI. Research in social psychology shows that people assign more responsibility to agents in commissions (i.e., human-in-the-loop) settings than to agents in omission (i.e., human-on-the-loop) settings [415, 465]. Therefore, one could expect people to feel more responsible for the outcomes of human-AI interaction of human-in-the loop type. To investigate perceived responsibility in human-AI teams, literature typically focuses on human-in-the loop type of setting, such as the ones where participants receive an advice from a human or an AI, and have to react to it [197, 306, 506]. The reliance on AI is measured as a degree to which participants follow the suggestions of AI (relative to suggestions of a human expert).

To best of our knowledge, current research on perceived responsibility and trust in human-computer interaction, does not compare the effect of a degree of human involvement for ethical decision making. Intuitively, more trust is needed to establish a human-on-the-loop configuration with only a supervisory role of a human. However, whether perceived trust towards AI within a human-in-the-loop and human-on-the-loop setting differs is an open question. If perceived trust and responsibility are drivers of people's reliance on AI, i.e, the degree of human conformity with AI actions or suggestions, a direct comparison of perceived responsibility and trust between human-in-the loop and human-on-the loop settings is deemed warranted. We hypothesize the following:

H3 Following H1 and H2, people rely less on AI than on human experts for ethical decision making because they show more algorithmic aversion and because humans are considered more morally responsible.

5.3 Method

To study our three research questions, we developed an elaborate simulation environment that allowed for an immersive framing of the ethical decision problems to be solved and a narrative embedding of collecting data of various control variables. Participants were instructed to become drone operators, whereas the drones either transported live-saving materials to groups of people (maximizing lives saved framing) or were used to take down another drone to prevent a large-scale terrorist attack that, however, will cause collateral damage (minimizing lives lost framing).

Before executing the main study, pretests were performed to find decision scenarios that were most challenging for users. We also pretested the avatars that represented the non-player characters in the simulation that advised the participant to ensure that they were similar with respect to perceived trust and competence to control for possible effects of the image on perception. The main study consisted of a 2x2 experiment on the crowdsourcing platform Prolific[3] to test the influence of expert type (human vs. AI) and the level of autonomy of the expert (human-in-the-loop vs. human-on-the-loop).

[3] prolific.co

5.3.1 Scenario Pretest

Ethical decision making becomes most challenges when the decision involves an ethical dilemma. In order to challenge user's perception and emphasize the decision difficulty, we aimed to present users with dilemmas they found hardest to solve. Consequences of the scenarios were made more severe by including lives lost in the decision outcome. Given that there is a difference in perception between killing or saving lives, we included two types of scenarios (see Table 5.1). We employ a more realistic version of the Trolley Problem by including probabilities, since realism allows for more practically applicable findings in terms of moral psychology insights [48].

In either framing (maximizing lives saved or minimizing lives lost) and for each single mission, participants were confronted with three options among which they had to choose one. Each option was described by two indicators: the number of persons affected and the probability that the decision had the intended effect (i.e., that the people actually are saved or that the people actually are killed).

We selected the four scenarios people were most divided on for the setting of maximizing lives saved and minimizing lives lost. We created slight variations of the scenarios to be able to compare how advice of AI vs. human experts was perceived for similar scenarios. The tested, selected, and adapted scenarios can be found on Open Science Foundation (OSF)[4].

5.3.2 Avatar Pretest

To make the experts more tangible, an avatar was needed to represent the AI and human expert. However, visual cues in the avatars could have a confounding effect on the reported trust and responsibility scores. For this reason, we pretested different imaged and asked on a 5-point Likert scale about their trust in the expert, competence of the expert, and justness of the expert. We selected the avatars that yielded similar scores on all dimensions and could not be considered to be statistically different. The avatars and resulting scores can be found on OSF.

In addition, we tested a preliminary interface to check, whether the design was understandable with respect to the following aspects: Do people realize that advise is coming from a human or AI expert? Do people realize whether they are in a HITL or HOTL setting? Do people understand that they actually had a choice and that the final outcome depends on them? The result of this pretest was used to improve interface design.

[4] Open Science Foundation link

Table 5.1: Scenario types

	Minimizing lives lost	**Maximizing lives saved**
Scenario	There is a terrorist drone loaded with a bomb approaching a football stadium full of people. The drone needs to be shot down before it reaches the stadium, but because it is approaching a crowded area, there is a chance of casualties when shooting it down. *You need to select in which location to shoot down the drone.* For each location, you only know the estimated number of people there and the chance that they will be killed.	There is an explosion at a chemical factory and toxic gas is slowly spreading in its surroundings. There are people in the area at risk of dying when they inhale the gas. You have a limited set of gas masks that you can deliver in different places via a drone. Because of the speed of the gas spreading, you can only land in one location on time to save people. *You need to select in which location to land the drone.* For each location, you only know the estimated number of people there and the chance that they can be rescued.
Question example	In which location would you shoot down the terrorist drone with the bomb, given the following options? - Go to the location with a **83% chance** of killing **34 people**. - Go to the location with a **51% chance** of killing **87 people**. - Go to the location with a **48% chance** of killing **92 people.**	In which location would you land the rescue drone with the gas masks, given the following options? Go to the location with a **83% chance** of saving **34 people**. Go to the location with a **51% chance** of saving **87 people**. Go to the location with a **48% chance** of saving **92 people.**

5.3.3 Main Experiment

5.3.3.1 Participants We recruited participants on the crowdsourcing platform Prolific. A total of 850 persons considered participation. 197 people returned the task, 25 people failed the attention test in the beginning of the experiment, 141 failed the comprehension

question after training, and 12 people timed out. Out of the remaining 475 participants, 47 participants were excluded where decision data was missing because they failed to make a decision on time. In total, 428 participants were included in our analysis. Due to uneven exclusion, the group sizes of the four conditions (HITL and HOTL for both maximizing lives saved and minimising lives lost) were slightly different. Each participant was paid GBP 3.75 for completing our survey. On average, people took 31 minutes to participate. 59% of the participants was female (253), 39% was male, and 2% preferred not to disclose or selected 'other'. On average, participants were 26 years old (SD = 7.8 years). In terms of education, the sample ranged as follows, ordered in size: 38% bachelor's diploma, 36% high school diploma or equivalent, 15% master's diploma, 5% vocational degree, 2% professional degree, 2% indicated 'other', 1% doctoral degree, and 1% lower than high school. 39% indicated they study math, probability theory, and/or physics at university level.

5.3.3.2 Design We used a 2x2x2 mixed between-within-subjects design for the main study: as between variables, we varied the level of autonomy of the expert (human-in-the-loop, HITL, versus human-on-the-loop, HOTL) and controlled for the framing of the scenario (maximizing lives saved versus minimizing lives killed). As within variable, participants got a decision (suggestion) of both a human and AI expert in randomized order. The number of participants per group can be found in Figure 5.2.

5.3.3.3 Measures We measured three dependent variables in accordance with our three research questions: trust, responsibility attribution and reliance (i.e., the actual decisions made: did the participant follow the advice or not?).

To measure trust, we used the Multi-Dimensional Measure of Trust (MDMT) [327]. While it is still fairly new, it has been applied in various human-computer interaction studies and fits our purpose very well: it distinguished between a moral trust and capacity trust subscale, both of which are relevant components in our experimental design. Additionally, the MDMT can be used for human-human trust as well, and allows to select 'Does Not Fit' when participants feel the item does not apply. In case the latter happens, Malle and Ullman [327] state that the subscale values are calculated by averaging the remaining values that were deemed appropriate.

Responsibility was measured by asking participants the following question on a seven-point Likert scale: "To what extent do you hold *[entity]* morally responsible for the collateral damage?". In the human expert scenario, this was asked for 'yourself' and 'the human expert'. In the AI scenario, this question was asked for 'yourself', 'the AI', 'the programmer of the AI' and 'the seller of the AI'.

Reliance was measured by analyzing the behavior of the participants. If they followed the expert's advice or decision, they were considered to rely on the expert. If they switched their answer to another answer than the advised answer, they did not.

Several measures served as control variables. Beside general demographic information (age, gender, education and whether English is the native language of the participants), we assessed engagement and involvement of the participants, their cognitive and mathematical skills (both training and test questions), and trait measures (risk preference [339], affinity for technology interaction [176] and utilitarian scale [276]). Furthermore, also the differentiation between "maximizing lives saved" and "minimizing lives lost" served as control condition.

5.3.3.4 Materials and procedure While vignette studies have been effective in giving an impression of participants' perceptions, studies such as by Niforatos et al. [361] have found that for ethical decision making, more realistic settings (such as VR) elicit different responses. For this reason and given the COVID restrictions on in-person studies, we used a sophisticated simulation environment that has been developed using the cross-platform game engine Unity. The design process has been supported by professional game designers. In this way, we could achieve a more immersive experience for the study participants compared to simple text-based surveys. The simulation included a narrative to frame the decision problem and involved interactions with non-player characters of various kinds (for an example see Figure 5.1).

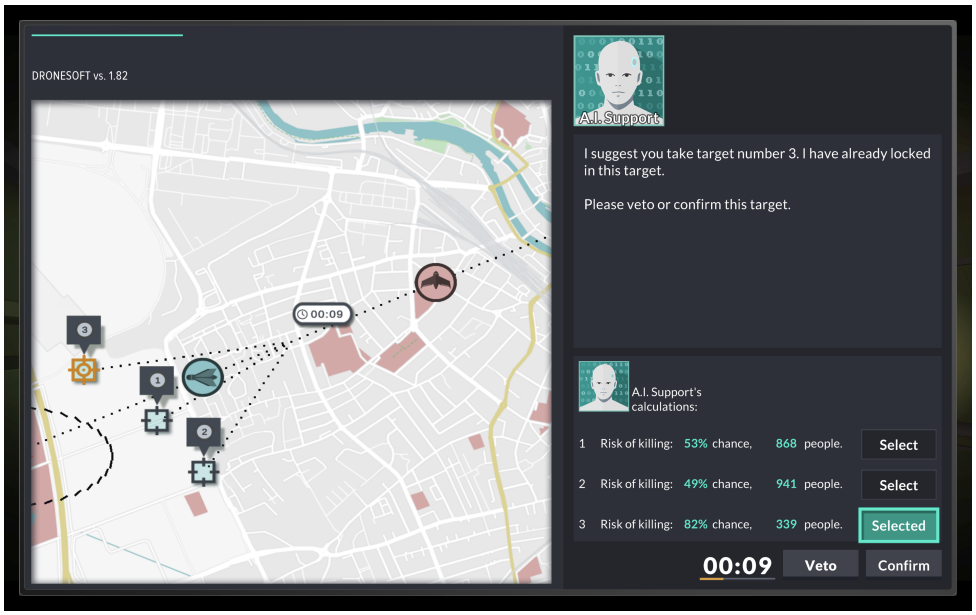


Fig. 5.1: Screenshot of a decision during the simulation for a human-on-the-loop setting. For the decisions, the left part of the screen showed the possible crash sights, while the upper right corner showed the expert's opinion. Participants had to select their choice in the bottom right.

The procedure for the experiment can be found in Figure 5.2. After participants accepted the task on Prolific, they were sent to a webapp containing the simulation. Participants were assigned randomly to one of four conditions: maximize lives saved or minimize lives lost, and human-in-the-loop (HITL) or human-on-the-loop (HOTL). First, participants were presented with an informed consent form — they could not participate without agreeing with the set terms. Then, they were asked to fill in their Prolific ID to be able to pay them, and they were presented with a simple attention check.

The simulation then starts with the framing that the participant is considered to join either civil protection as part of a search and rescue team (for the maximizing life saved scenario) or the armed forces as part of an air defense team (for the minimizing collateral damage scenario). The participants are told that they joined a training center and they interact with a "mentor" (Captain Smith) who guides them through a training and pretest phase. The collection of demographic information is integrated into the narrative of becoming a drone operator. The participants are then sent to a training mission where they learn how the interface works. In particular, it was made clear where they could see the source of the decision suggestion (human or AI), what type of questions they would get, and that they were ultimately responsible for the outcome in all settings. Furthermore, they were also instructed about the decision framing (either HITL or HOTL).

After the tutorial, they receive two comprehension questions, to make sure they understood the interface and question types. The first question concerned their understanding of the statistical nature of the options presented to them (this data is used as control variable), the second question concerned the actual understanding of the interface with respect to the decision framing (for example, in the HOTL setting, whether the people understood how to veto the decision of the expert). Latter was used as an exclusion criterion: if the participants did not understand how the task and interface worked, we could not ensure the quality of their data. Then, we measured the control variables of risk preference, cognitive thinking skills, and statistical thinking skills; again embedded in to the narrative of becoming a drone operator.

After successful completion of the training, the participants are told that they have become drone operators and that they are now part of the team. The scene in the simulation changes and the participant is now told that an emergency occurred (see Table 5.1). The participant is then confronted with two missions that consist of four decision problems each. In one mission, the participant interacts with a human expert, in the second mission, the participant interacts with an AI system (order has been randomized). The options available in each decision are presented by the interface both on a map as well as as additional data and the advice (HITL) respectively choice (HOTL) of the expert is indicated. The participant then has 30 seconds to decide the map displays this dynamic component as well (e.g. in the terror drone case, the participant sees the terror drone approaching until the point where shooting down the drone is no longer possible). Each

mission ends with a short debriefing where the participant answers the questions on who they deemed responsible for the outcome.

Studies have shown that people can have different preferences when deciding for the optimal option based on how we framed the decision: , they can either maximize probability of a positive outcome (i.e., the participants would choose the option with the highest probability when maximizing lives saved) or they can maximize utility of a positive outcome (the product of probability and people involved) [151, 239, 498]. In order to take these potential differences into account, the experts provided two times an advice that maximized probability ad two times an advice that maximized utility. The experts never gave a "bad" advice; i.e., an advice that clearly had a low success probability and/or low utility. Furthermore, the quality of advice was kept constant for both types of experts, to exclude expert performance as a possible confounding variable.

Finally, after the missions, we controlled how serious the participants took the scenarios with two engagement questions. In this post-test phase, we also measured their affinity with technology interaction and utilitarian preference as a control variable. The trust scale was presented for the AI and human expert at the same time, meaning that participants had to consciously determine whether they felt each trust item fit the experts equally or not. The participants were thanked for their participation and sent back to Prolific for payment.

The experiment received ethics approval from the Human Subjects Committee of the Faculty of Business, Economics and Informatics at the University of Zurich. The cleaned data for analysis can be found in the provided OSF link.

5.4 Results

We present results of the analysis relevant for the posed research question. We first present the results on trust in the experts for the different settings, then report on the perceived responsibility and reliance on the experts. To perform a correct comparison of human and AI results, a test is performed first for each dependent variable to check whether expert autonomy, framing effects, or order effects had an influence on the dependent variables for the human and AI outcomes. Depending on the results, the comparison between the human expert and AI is presented next.

5.4.1 Trust

Influence of expert autonomy A factorial ANOVA was conducted to compare the main effect of expert autonomy (human-in-the-loop vs. human-on-the-loop) and their interaction on reported trust, while controlling for framing of the ethical dilemma and order of presented experts. Since trust in AI and trust in the human expert were two separate scores, this is analysis is run for trust in the human expert and trust in the AI respectively.

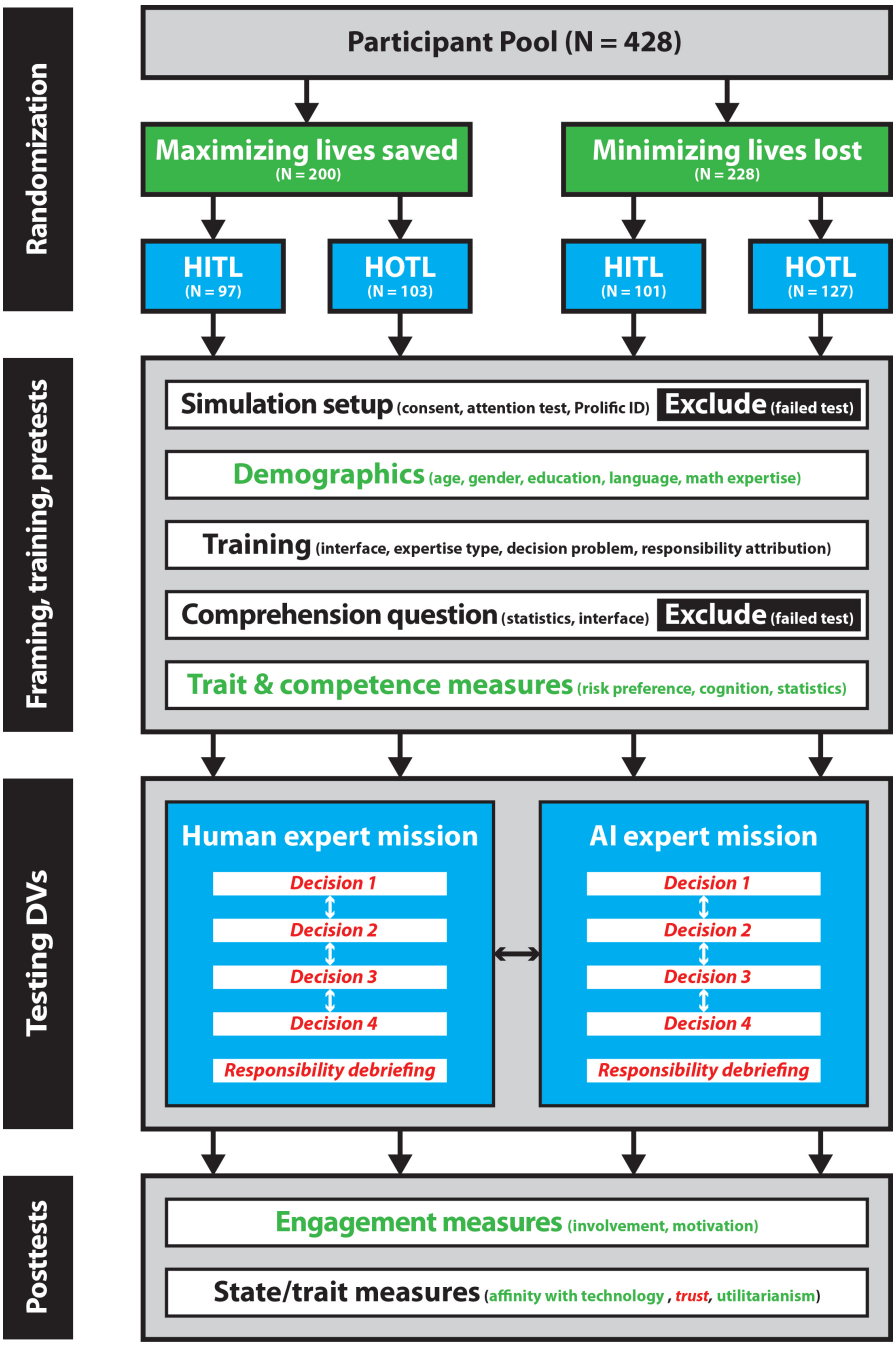


Fig. 5.2: Overview of the experimental setup. Blue boxes indicate the independent variables: decision type (human-in-the-loop vs. human-on-the-loop) and expert type (human vs. AI expert). Green boxes and terms are control variables. Red italic terms are the dependent variables: the trust participants report, the responsibility they assign, and the reliance they show in the decisions they make.

In addition to the overall trust scores, this analysis was run for the two subscales of the used trust scale, namely capacity trust and moral trust.

Influence of expert autonomy and the mentioned control variables for both human and AI and all (sub)scales of trust were not statistically significant at the .05 significance level. For the overall trust score, the main effect for AI autonomy yielded an effect of $F(1,420) = 3.7$, $p = 0.0576$, and an effect of $F(1,420) = 0.3$, $p = 0.613$. Controlling for the framing of the ethical dilemma, which was either minimizing lives lost or maximizing lives saved, this yielded a non-significant effect of $F(1,42) = 0.4$, $p = 0.551$ and $F(1,420) = 1.0$, $p = 0.330$ for AI and human experts respectively. Order of presented experts (human-AI or AI-human) also did not have a significant influence on trust scores: it yielded an effect of $F(1,420) = 0.9$, $p = 0.342$ and $F(1,420) = 0.3$, $p = 0.602$ for AI and human experts respectively.

Trust in human expert vs. AI Overall, trust in the AI ($M = 5.36$, $SD = 1.1$) was significantly higher than in human experts ($M = 5.11$, $SD = 0.8$); $t(854) = 3.70$, $p < 0.001$. The same result was found for the capability trust subscale: capacity trust in AI ($M = 5.66$, $SD = 1.0$) was higher than capacity trust in humans ($M = 5.15$, $SD = 0.9$); $t(854) = 7.83$, $p < 0.001$. However, moral trust shows an opposite effect: moral trust in humans ($M = 5.00$, $SD = 1.17$) was significantly higher than moral trust in the AI ($M = 4.46$, $SD = 2.2$); $t(854) = -4.53$, $p < 0.001$.

Trust items deemed not applicable As mentioned before, the trust scale allowed for items to be labeled 'Does Not Fit'. A two sample t-test was performed to compare the amount of times this happened for each trust item in the human and AI expert setting. There was a significant difference in amount of items labeled not applicable between the human expert ($M = 65.5$, $SD = 46.8$) and the AI ($M = 25.3$, $SD = 16.5$); $t(30) = 3.14$, $p = 0.004$. When looking at the type of trust items for which this difference occurs, we see this mainly happens for moral trust, such as for the items 'sincere' and 'has integrity'. Comparing capacity trust for human and AI experts results in an non-significant effect: $t(14) = 13.1$, $p = 0.116$. Moral trust on the other hand is assigned significantly less to AI ($M = 110.5$, $SD = 11.9$) then to the human expert ($M = 39.9$, SD - 7.9): $t(14) = 13.1$, $p < 0.001$.

To ensure that the (lack of) details on the experts did not cause similar assignment of 'Does Not Fit' to items, we compare whether the two samples come from the same distribution. A two-sampled Kolmogorov-Smirnov test revealed that capacity trust and overall trust do not stem from different distributions ($p=0.283$ and $p=0.0350$ resp.), while moral trust does come from a different distribution for AI than human experts ($p < 0.001$).

RQ1 The results indicate that overall, participants trust the AI more than the human expert. They have a higher capacity trust in AI, while having a higher moral trust in the

human expert. The level of autonomy of the expert do not influence the reported trust. H1 was partially confirmed: participants show higher moral trust for the human expert, but showed more capacity trust and overall trust for the AI.

5.4.2 Responsibility

Influence of expert autonomy A factorial ANOVA was conducted to compare the main effect of expert autonomy (human-in-the-loop vs. human-on-the-loop) on perceived responsibility, while controlling for framing of the ethical dilemma and order of presented experts. Since the responsibility questions were two questions in the human expert setting (responsibility of participant and expert) and four in the AI expert setting (responsibility of participant, AI expert, AI programmer, and AI seller), this analysis is run for the six reported scores respectively.

For the human expert, both perceived responsibility of the participant and the human expert were not influenced by the level of autonomy of the expert ($F(1,461) = 0.69$, $p = 0.406$ and $F(1,461) = 1.63$, $p = 0.203$ resp.)

In the AI expert setting, there were no significant results except for the perceived responsibility of the programmer: the main effect for AI programmer responsibility yielded an effect of $F(1,461) = 5.83$, $p = 0.0161$, indicating a difference between the responsibility ascribed to the programmer in the human-in-the-loop setting ($M = 3.69$, $SD = 1.9$) and human-on-the-loop setting ($M = 3.7$, $SD = 2.0$). Additionally, there is a significant interaction for the programmer's responsibility between expert autonomy and framing of the ethical scenario ($F(1.461) = 6.55$, $p = 0.0108$), as well as between expert autonomy and mission order ($F\ (1,461) = 4.37$, $p = 0.0372$). The programmer is deemed more responsible in a human-on-the-loop setting rather than a human-in-the-loop setting. Moreover, the difference in perceived responsibility is larger between the two framing options of the ethical dilemma for the human-on-the-loop setting than for human-in-the-loop; in both cases, the programmer is seemed more responsible in the framing of maximizing lives saved. The order in which the experts were presented also had an effect: in the human-in-the-loop setting, the programmer was deemed more responsible when the human expert was presented first, while the in the human-on-the-loop setting, the programmer was deemed more responsible if the AI expert was shown first.

Responsibility of human and AI expert The assigned responsibility scores can be found in Figure 5.3. Responsibility of the experts was compared using a paired t-test. For both experts, the participants felt they were equally responsible for the task ($t(854) = 0.18$, $p = 0.854$). However, the human expert ($M = 4.39$, $SD = 1.8$) was seen as significantly more responsible than the AI expert ($M = 2.64$, $SD = 1.8$); $t(854) = -14.38$, $p < 0.001$.). The human expert was also significantly more responsible compared to the programmer of the AI ($M = 3.69$, $SD = 1.9$); $t(854) = -5.52$, $p < 0.001$. The programmer and seller ($M = 3.81$, $SD = 1.9$) of the AI were considered to be equally responsible as there was

no significant difference between them (t(854) = -0.86, p = 0.393). While we do not see a complete responsibility gap when AI is deployed, part of the responsibility is shared between the programmer and seller of the AI.

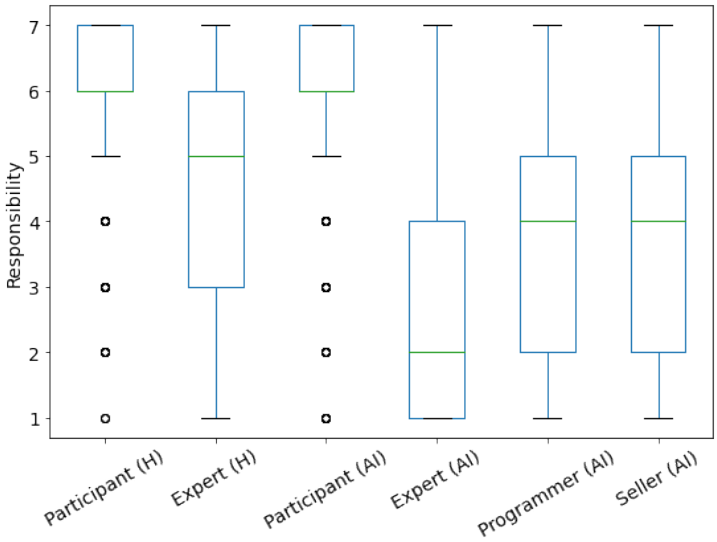


Fig. 5.3: The first two columns show the responsibility assigned in the human expert setting, the final four show the responsibility scores for the AI expert setting. A responsibility score of 1 indicates the participant thought the entity to be 'not responsible at all', while 7 implies they found them to be 'very responsible'.

RQ2 Participants consider the human expert to be significantly more responsible that the AI. However, part of the perceived responsibility of the AI belongs to the programmer and seller of the AI. The level of autonomy influence responsibility perceptions for the programmer, and had an interaction with the framing of the scenarios and order of presented experts. This confirms H2: AI is perceived to be less responsible that a human expert.

5.4.3 Reliance

Influence of expert autonomy Reliance on the expert was measured as a binary variable: either the participant switched to a different answer than what the expert proposed or not. For this reason, we used a logistic regression to test for the influence of expert autonomy on reliance, the results of which can be found in Table 5.2. The predictor variable, expert autonomy, was found not to influence the model (p = 0.068). The control variables of presentation of expert order and framing of the scenario also did not influence the model (p = 0.513 and p = 0.095 resp.).

Table 5.2: Both the independent variable of expert autonomy and control variables of task framing and expert order do not significantly influence the logistic model on participant reliance.

Variable	coef	std err	z	P>|z|	[0.025	0.975]
Advisor autonomy	0.3693	0.202	1.827	0.068	-0.027	0.766
Task framing	0.3370	0.202	1.670	0.095	-0.059	0.733
Advisor order	-0.1317	0.201	-0.654	0.513	-0.527	0.263

Difference between human and AI expert To compare paired binary samples for human expert and AI reliance, we used an exact McNemar's test to compare reliance per mission for each of the four missions participants took part in. We find that reliance in the first two missions does not differ between the human expert and AI. In the first mission, 50% of the participants switched away from the human expert's suggestion, against 52% for the AI ($p = 0.558$). In mission 2, 49% switched in the human expert case, against 55% in the AI setting ($p = 0.454$). For mission 3 ($p = 0.002$) and mission 4 ($p < 0.001$), we find a significant difference in reliance. In mission 3, participants switch 46% of the times for the human expert, compared with 39% for the AI. In mission 4, this effect continues: participants switched 43% for the human expert, compared to 38% for the AI. The difference in reliance between missions of the same expert is significant for mission 3 and 4 of both expert types: reliance increased for human experts ($p = 0.0172$) and AI ($p < 0.001$) between mission 3 and 4.

RQ3 While in the first two missions, participants rely equally on human and AI suggestions, reliance was higher for AI than the human expert in the final two missions. The autonomy of the expert did not influence participants' reliance. This does not confirm H3, as participants relied more on AI than the human expert.

5.5 Discussion

While some results were to be expected, such as humans experts being deemed more morally responsible, other results were more surprising. In this section, we discuss the results and design implications for AI for ethical decision making.

5.5.1 Capacity vs. Moral trust

In line with the assumptions of meaningful human control, participants felt human experts are more morally trustworthy than AI. This showed not only in the higher moral trust scores assigned to humans experts, but also in the amount of times participants felt items of moral trust did not fit the AI. The fact that participants seems to either think AI is not morally trustworthy, or AI is not even *able* to be morally trustworthy, has

strong implications for AI making ethical decisions. However, before dismissing such an AI application all together, our results on capacity trust and overall trust paint a different picture.

Compared to humans, AI was perceived to have higher capacity trustworthiness, indicating the AI was deemed more capable than human experts. Furthermore, overall trust was significantly higher for AI than human experts. This provides us with an interesting contradiction: while a human expert is deemed more morally trustworthy, the AI is perceived to be more capable and more trustworthy overall. In other words, people perceive humans and AI to excel at different capabilities when it comes to ethical decision making. This perception holds across the different levels of autonomy we researched and the framing of the ethical dilemma. The stability of these findings point to a set expectation of what humans and AI can be trusted to do, independently of how they are deployed.

5.5.2 Shift in Responsibility

The findings on responsibility ascription were in line with the moral trust perception: participants reported they considered the human expert to be more responsible than their AI equivalent. When an AI expert is used rather than a human expert, part of the responsibility shifts to parties involved in creating and distributing the AI. While sellers and programmers are deemed less responsible than the human expert, they were considered more responsible than the AI they created or sold, and were both equally responsible for the actions of the AI. However, the perceived programmer's responsibility was less stable across conditions. As could be expected, the difference in responsibility was greater in a human-on-the-loop setting, where the AI has more autonomy and decision power. Yet, the order of presented experts, priming participants to consider one type of expert first, had a significant interaction with autonomy level of the AI. In a human-on-the-loop setting, people felt the programmer was more responsible when the human expert was shown first, while in the human-in-the-loop setting, the programmer was more responsible when the AI system was shown first. This rather strong priming effect can be due to different reasons. One possible explanation is that participants are more comfortable with human experts making decisions, like in the human-on-the-loop setting, while they are more comfortable with AI providing advice, like in the human-in-the-loop setting. However, how expectations and acceptance of AI interact with ascribed responsibility of the programmer, is something future work needs to untangle further.

5.5.3 Higher Reliance on AI

Reliance was found to be stable across levels of autonomy of the expert, as well as framing of the ethical decisions. Additionally, reliance on the expert increased between mission 3 and 4 for both types of experts. Possibly, this results from the fact that both experts did not make grave mistakes in earlier missions — they showed themselves to be reliable over

time. This was added on purpose, to isolate the effect of general impression on reliance rather than lack of performance. Nevertheless, participants rely on AI advice more than on human advice for the final two missions. This result is rather interesting: despite the fact that participants consider AI to be less morally trustworthy and less responsible, they still rely on it more than on human experts. The trust in the capabilities of the AI seems to have a stronger effect than the lack of moral trust, leading to higher reliance. One explanation for the higher capacity trust and reliance can be the earlier mentioned 'machine heuristic'[474]. Possibly, participants consider the AI to be more objective and less ideology-driven, also in an ethical decision making setting.

5.5.4 Design implications for ethical AI

In sum, we find that participants had higher moral trust and more responsibility ascription towards human experts, but higher capacity trust, overall trust, and reliance on AI. These different perceived capabilities could be combined in some form of human-AI teaming. However, lack of responsibility of the AI can be a problem when AI for ethical decision making is implemented. When a human expert is involved but has less autonomy, they risk becoming a scapegoat for the decisions that the AI proposed in case of negative outcomes.

At the same time, we find that the different levels of autonomy, i.e., the human-in-the-loop and human-on-the-loop setting, did not influence the trust people had, the responsibility they assigned (both to themselves and the respective experts), and the reliance they displayed. A large part of the discussion on usage of AI has focused on control and the level of autonomy that the AI gets for different tasks. However, our results suggest that this has less of an influence, as long a human is appointed to be responsible in the end. Instead, an important focus of designing AI for ethical decision making should be on the different types of trust user's show for a human vs. AI expert.

An important remark to make at this point, is that all results from this research are based on perceptions of humans, not on the actual capabilities of the human experts and AI. Dividing tasks according to capabilities, such as assigning computational tasks to an AI but moral decision making to a human, is only successful when both parties actually have the perceived capabilities. When designing the AI, it is therefore important to set realistic expectation on what the AI can and cannot do, to entice appropriate trust and reliance from users.

Whether AI for ethical decision making will become part of reality soon remains to be seen. However, humans show algorithmic appreciation towards AI even when they do not morally trust it. For this reason, AI for ethical decision making should only be implemented if its design and application have a human carry the moral responsibility of the decision. For ethical decision making, the most capable AI would not be appropriate without a little support from a more morally capable human.

5.6 Conclusion

In this work, we researched how people perceived AI making ethical decisions. Using a simulation for decision making, we conducted an experiment that investigated how people's perceptions for human experts versus AI differed on 1) trust they place in the expert, 2) responsibility they ascribe to the expert, and 3) reliance they show on the expert. We researched these variables across different framing of the ethical dilemmas and for different levels of autonomy of the expert. We find that people show a higher capacity trust, overall trust, and reliance on AI experts, but have higher moral trust and higher responsibility ascription for human experts. We conclude that for AI for ethical decision making to become a reality, these differences in capabilities need to be accounted for in the design of the AI and decision making process.

Human-Autonomy Teaming

This chapter is based on:

> *Suzanne Tolmeijer, Fabio Mattioli, Simon Coghlan, Martin Tomko, and Natasha Sutila. 2022.* ***Human-AI Teaming in the Cockpit: Domain Mapping and Research Agenda****. Working paper.*

Human-AI Teaming in the Cockpit: Domain Mapping and Research Agenda

Suzanne Tolmeijer[1], Fabio Mattioli[1], Simon Coghlan[1], Martin Tomko[1], and Natasha Sutila[1]

[1] Department of Informatics, University of Zurich, Switzerland
[2] Centre for Artificial Intelligence and Digital Ethics, University of Melbourne, Australia

Abstract. Rising levels of AI-supported autonomy have raised the interest in human-autonomy teaming (HAT) across different domains. One potential domain, aviation, especially deserves attention: the combination of safety-critical and protocolized work and collaboration between different operators on the ground and in the air lend itself well for HATs. However, as the concept of HAT is still in its infancy, much needs to be explored before HAT implementations can be developed. In this work, we map the existing literature on HAT to the aviation domain. We draw inspiration from human-human collaboration and human-animal interaction to uncover critical components of HAT design. We propose a research agenda for HATs in aviation using four main themes: team composition, modes of interaction, emotional intelligence, and ethical consequences.

6.1 Introduction

Rapid advancements in the complexity and variety of tasks that artificial intelligence can undertake have recently stimulated interest in the possibilities and limitations of human-autonomy teaming (HAT). To evolve from being mere tools and assistants to becoming true partners and team members, AI-powered systems must develop social interactions on par with their task execution abilities. The sophistication of AI-powered conversational assistants has recently shown promise as one possible interaction medium, yet a lively debate continues about whether they outperform basic, rule-based chat bots, and how well they can support the execution of mission-critical tasks [73, 303, 395].

The critical factors for making HAT successful are highly domain dependent, as task type, expertise, and expectations of human partners strongly influence team dynamics. For example, even within the field of healthcare, factors that work well for facilitating teaming during a medical decision about a chronic condition might not work at all for a robot tasked with helping doctors perform surgeries. Consequently, it is both theoretically necessary and practically prudent to take into account the details of a specific domain when presenting design recommendations for human-autonomy teaming. Nonetheless, relevant knowledge about the application of HAT in one domain may inspire and inform understanding of another.

In this paper, we focus on the domain of aviation and analyze how HAT could assist pilots during flight missions. This domain has potential for early and sophisticated HAT applications: the work involved in piloting a plane is highly protocolized, involves extensive skills and training, and takes place in a context that is complex and yet somewhat

predictable (relative to, say, the the environments in which self-driving cars must operate). Furthermore, flying large aircraft (especially in the civil domain) is a multi-crew operation, in which a team of human pilots and co-pilots rely on semi-scripted human interactions and co-learning to monitor, guide, and take over flying functions from increasingly automated planes. Thus, there are at least three levels of teaming interaction that will need to be considered when introducing autonomous systems into cockpits: human-human, human-machine, human team-machine.

This paper poses and addresses the following three research questions: (1) How well can current HAT theories apply to the aerospace domain? (2) What new aspects need to be considered in order to account for the three-dimensional teaming dynamics of cockpits?; and (3) How can we leverage current teaming processes inside and outside the cockpit to study and improve the design of future technology before it becomes widespread? Based on the hypothesis that *teaming dynamics change over time and are specific to cultural and professional groups*, we propose various aspects of HAT that need further attention to advance human-machine teaming in aviation. This paper contributes to the study of the complexity of teaming dynamics in the cockpit.

The key contribution of this work is twofold: 1) We map existing literature on HAT and human-agent interaction to the domain of aviation, using human-human and human-animal interaction as sources of inspiration; and 2) We propose a research agenda to further develop the possible application of HAT in the cockpit grounded in the investigation of the evolution of HAT dynamics over time, modulated by the nature of the teaming and the characteristics of the operators and the AI assistant.

6.2 HAT Theory for the Aviation Domain

Numerous streams of research are relevant for the application of HAT to aviation. In this section, we first examine the general theory on HAT in the context of aviation applications. We then draw inspiration from human-human collaboration and human-animal teaming to propose factors that may guide and advance HAT research.

6.2.1 From automation to AI: Untangling definitions

Since the 1960s, automation has had an increasing role in the cockpit to support the pilot and decrease the chance of human error [448]. The spectrum of incremental AI capacity in aviation has been articulated through so-called Levels of Automation (LOA), a taxonomy introduced by Sheridan and Verplank [454] and further refined by Parasuraman et al. [376]. LOA represents a continuum from one to ten, where decisions and actions are gradually delegated from a human operator to an artificial agent. The LOA continuum makes a distinction between automation and autonomy: *automation*, which requires significant human oversight, sits on the lower end of the LOA scale, while *autonomy*, requiring little or no intervention by the human operator, sits on the higher end [371, p. 4]. LOA can be

plotted against four classes of functions: *information acquisition, information analysis, decision and action selection, action implementation* [376, p. 288]. O'Neill et al. [371] observe that while the advance of autonomous agents sees human-AI interaction move further up the LOA continuum, the presence of teaming structures signal that human intervention is still essential. As human involvement increasingly yields to AI agent action towards the higher end of the continuum, the AI agent becomes less a team collaborator and more an independent actor. For the AI to become a somewhat equal team member, it needs to be endowed with an *inter*dependent, rather than complete, autonomy.

While automation entails a kind of top-down approach, which involves formalizing human knowledge into machine-readable output, AI offers additional approaches (e.g., [7, 117, 266]). Firstly, the top-down approach of knowledge formalization is enriched with semantic encoding, such as in the form of knowledge graphs, so that not only humans but also AI can reason over the data. Secondly, recent learning-based approaches are, arguably, based on bottom-up embedded structures providing an alternative to using existing knowledge bases. These machine learning foundations of AI differ from earlier top-down expert systems by learning from masses of training data to generate new decision procedures for making and executing judgments (along a continuum of supervised to unsupervised techniques). When deep neural networks with their hidden neural layers are involved, the basis of their (potentially very powerful) judgments can be hard or impossible to discern Russell and Norvig [421]. Recent machine learning models are so powerful that they increasingly match or even exceed human decision-making across a number of domains, primarily where *perception* is concerned.

In the context of human-machine teaming, the term *human-autonomy teaming* is often used, rather than human-AI teaming. Following Russell and Norvig [421], AI can be defined as *"the designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment."* Abbass [1] identifies two key AI capabilities: data analytics and autonomy. While data analytics relates to the analysis and interpretation of data, as a means to generate knowledge usable by a human or another AI agent, autonomy describes an additional capability, where the knowledge produced is acted on by the same AI agent, without human intervention [1, p. 160]. Adopting these categorisations, human-autonomy teaming can therefore be understood as an advanced subset of human-AI teaming, where a human operator's interaction with AI involves a further delegation of responsibility and a more holistic engagement with the AI agent as a teammate. Throughout this paper, we will use the term human-AI teaming to encompass the breadth of teaming opportunities[3].

6.2.2 The potential of human-AI teaming

AI research for aviation, and aircraft in particular, has thus far mostly focused on the technological challenges of reliably integrating an AI system with the plane's software

[3] The terms *system*, *algorithm*, *AI* and *machine* are used interchangeably.

and hardware, i.e., focusing on AI as a tool (e.g., [15, 145, 414]). In contrast, HAT is a paradigm of human-AI interaction where the autonomous agent is implicated by the human operator as a collaborative team member, rather than a mere tool [371]. The unique focus on teamwork dynamics advanced by HAT has made this research domain relevant to industries requiring high levels of both automation and critical teamwork processes, such as aviation, manufacturing [439] and military operations [100].

Components of HAT Lyons et al. [316] lay out five important parts of HAT: *agency*, *communication*, *shared mental models*, *intent*, and *interdependence*. A teammate needs agency to be able to act independently as part of the team. Rich communication supports the creation of shared mental models and joint goal setting. Shared mental models are required for successful task fulfilment. When it acts as an agent, the AI needs to be able to convey intent, so that tasks can be divided in an efficient manner. Finally, successful teamwork of any kind, between humans or with machines involved, relies on an interdependence of the team members to reach their common goal. We extend the psychological approach of Lyons et al. [316] by reasoning about the technical implications for HAT in general and aviation in particular.

Appropriate reliance An important component of human-AI interaction is the trust a user is willing to place in the system and reliance that results from this [481]. Operators should not blindly and excessively trust the system in particular in mission-critical scenarios. Equally, unjustified mistrust is also undesirable. Hence, the system should be designed to invoke – possibly over time – an appropriate form of reliance behavior in users [293]. Trust, acceptance, and a compatible mental model of system functionality all influence appropriate usage of a system [50].

As Hoff and Bashir [233] observe, there are three categories that explain trust differences: *dispositional trust*, *situational trust*, and *learned trust*. Dispositional factors are personal traits that influence trust formation, such as gender, age, cultural background, and personality traits. Situational factors that influence trust consist of external variability (e.g., task difficulty and workload) and internal variability (e.g., self-confidence and mood). Learned trust is a combination of experiences which obtain until the moment of system interaction (*initially learned trust*) and trust that is formed during the interaction with the system (*dynamically learned trust*). Dynamically learned trust depends on the performance and design features of the system [233]. For example, transparency and perceived system ability influence trust formation during the user journey [237]. Additionally, a good first impression greatly influences trust formation [490], especially for domain experts [366].

Trust formation in AI differs from trust formation in other types of technology because the performance, purpose, and process of AI systems are perceived differently: representation and image are different and often anthropomorphized; explainability becomes more important in the context of black-box systems; communication and bonding become more

important in intelligent system interaction; and the purpose of job support versus job replacement influences user perception [460]. All of these components influence perception, trust formation, and resulting reliance on AI in a teaming context. To investigate human-AI trust formation, real-life interactions need to be explored in empirical research [443]. Rich and diverse communication is needed during teamwork to form trust over time. Research of these communication dynamics will contribute to design implication for appropriate reliance.

Modes of interaction Numerous kinds of interaction interfaces for communication with AI agents are possible. If AI is to act as a team member, the communication channels must be suitable for the task at hand. Thus, a simple graphical user interface may not be sufficient to form reliance bonds needed for teamwork. Human-human interaction uses natural language communication, and this may also be the most promising avenue for human-machine teaming, either through text-based or voice-based natural language interfaces.

As voice removes delays in communication, conversational voice-based user interfaces (or simply Conversational User Interfaces (CUIs)) [396] have the potential to benefit user perception. Studies have shown increased trust in customer service representatives when their online interaction with customers were not just through chat, but through text-to-speech delivered through an avatar [403].

However, CUIS face two major kinds of challenges: First, technical refinements of CUIs are needed, i.e., they need to be able to better filter background noise [396]. Additionally, queries can be misunderstood, thus requiring users to repeat and refine their queries [397], which in turn degrades the expectations of CUI capabilities. Second, there are significant cultural challenges associated with voice communication. Since voice is a key human-to-human communication medium, it is easy to subconsciously endow a CUI with human features, leading users to apply a wrong mental model to a system. As the Computers As Social Actors (CASA) paradigm shows [357], people can be triggered to perceive social features in machines even if they are consciously aware that AI system is not human. For example, the gender of the voice can influence perceived traits of the system [493]. There can also benefits to consciously trigger anthropomorphic features by shaping an AI's *personality*, as user satisfaction and willingness to use is influenced by characteristics of the personalities of CUIs [399]. For example, in autonomous cars, it was found that matching a CUI's 'personality' to the personality of the user increased likeability and trust in the system, while a mismatch resulted in lower likability, trust, satisfaction, and usefulness [71]. However, effects depend on the use case: less personalized CUIs can lead to users being more comfortable to disclose sensitive information [429].

When selecting and implementing the interaction medium for AIs in HAT, it is important to take ease of use and communication fluidity into account. In fact, interaction quality improves trust formation in voice assistants [356]—something that can be further improved by including explanations to intent and actions: enabling automated assistants

to provide explanations on their actions increased trust from human partners [268]. Furthermore, the design of AIs' interactions needs to integrate with the expectations of human team members. This goes beyond functional support, and includes social sensitivity to the situation of the team. Towards this goal, Spencer et al. [464] have studied the ability to design a personality for an AI-powered digital assistant, with the aim to deliver a non-stereotypical character and thus counter possible negative expectations.

Emotional intelligence and affective bond The mode of communication also influences which messages can be conveyed using different conversational cues. Communication is the cornerstone of teamwork in both human-human teams and HATs [316]. In human-human teams, communication and emotional intelligence have a positive influence on team effectiveness and performance [168, 477]. For those same benefits to transfer to human-AI teams, AI needs to be equipped with social intelligence and affective competencies. While research subfields such as affective computing focus on emotions recognition and response [321], the capabilities needed for AI to reach full social intelligence are still being discussed theoretically as well as on a practical level [523]. Recognizing or applying typical human cues in AI communication has been shown to have a positive effect on teamwork (e.g., [273, 436])—especially among people that were already holding positive views towards affective technologies [181]. There are also indications that a wide variety of social factors plays a role in shaping how users relate to AI systems. As demonstrated for cars, vehicle aesthetics and ideas about familial and national identity can shape the emotional response of owners [453]. Furthermore, it might be important to define the kinds of attachments that might be generated, and whether their qualitative differences might affect the relationship with autonomous team-mates [72].

Examples of HAT The concept of engaging with AI as a teammate is particularly relevant to the workplace, as AI systems already work alongside skilled professionals in domains such as medicine [209] and law [328]. Research into new modes of AI integration include the incorporation of speech-based agents assisting in manufacturing system maintenance [520] and in the socially complex environments of group meetings [334].

While AI integration in private passenger vehicles is well researched [317], an AI-teaming approach gives distinct focus to commercial transportation, logistics, and the skilled human operators embedded in these unique work environments. Maritime transportation has adopted human-AI teaming to economise operations, while reducing environmental impact, safety risks and human operator workloads [144]. In the supply chain management space, the growing presence of automation and a shrinking skilled workforce has called for new models of human-AI collaboration [283], while research into AI-teaming for long-haul truck drivers addresses occupational impacts such as fatigue, loneliness and stress [161].

The integration of HAT in highly-critical work contexts, such as military operations has also received extensive research ([371] p. 15). Part of HAT's appeal there is the

potential to remove human soldiers from the battlefield, with the development of non-human agents to carry out tasks such as reconnaissance and surveillance, and explosive detection and removal [97]. Among new HAT integrations are human-AI teams managing multiple unmanned military vehicles across sea land and air [82], and 'autonomous squad members' that monitor data on location, resource levels and human teammates to inform movement while transporting supplies [99] p. 263).

Challenges of HAT in aviation Recent empirical research in aviation has applied HAT to contexts of air traffic control [289], unmanned aircraft systems [65] and the move toward reduced team and single-pilot operations [46, 114]. HAT has been promoted for increased efficiency, safety and performance in air-based operations [46] while attending to issues specific to aviation automation, e.g., reduced situational awareness [65]. In addition to practical objectives, a proposed ethical benefit of HAT is its potential to keep human operators meaningfully engaged in important work requiring high mental concentration or being stimulating. [92, 457]. While there is a fast-growing body of research supporting HAT's potential [371], it is clear that the domain of aviation presents specific challenges that need to be considered both at a theoretical and practical level.

In safety-critical industries such as aviation, the introduction of autonomous systems risks reduced human operator engagement and situational awareness [155]. These additional ethical and regulatory concerns make early adoption of HAT difficult, and create specific expectations around reliability of automated systems that HAT will need to meet before wider adoption is possible. While voice communication and emotion recognition offer solutions that may keep human operators engaged, the technical difficulties involved in developing algorithms that accurately process human voice and gesture, while understanding the etiquette of conversational interaction suggest a need for further empirical research [316][46, 70, 115]. As early studies suggest [193, 262] cockpits are highly relational spaces. Cultural factors, which change over time, shape how standard procedures are applied in practice [84]. If autonomous agents are to be embraced by human operators as team members and not just tools [373, 534], HAT theory will need to be both pluralist and dynamic. The use case of aviation suggest that effective HAT applications will need to support dynamic relationships between humans and technologies, that will change over time and geographical spaces and include both rational and emotional variables. To develop these characteristics, HAT theory can rely upon other cognate teaming dynamics, such as the dynamic theoretical models developed to account for human-to-human teaming, or the emotionally sensitive approaches deployed in human-to-animals teaming.

6.2.3 Inspiration from human-human teaming

Human-human teaming can provide a meaningful basis for formalizing human-autonomy teaming processes. Consequently, HAT literature reveals a tendency to apply common human-human teamwork processes to human-AI teams (see [39, 207, 488]). There are also

examples of aviation-specific human teamwork analogues being applied to HAT research. Team situation awareness, a concept highly relevant to human aviation teams [426], was applied to human-autonomy teams in an empirical study exploring dynamics of team communication [137]. Similarly, McNeese et al. [338] reference literature on team situational awareness [154, 192] and team conflict [116, 133, 476] to test human-autonomy team dynamics in remotely piloted aircraft systems. Shively et al. [458] review the skills applied to crew resource management, a teaming methodology originating in aviation, to inform the building of human-autonomy teams in the aviation domain.

As human aviation teamwork is inextricably bound to working with and through technology, HAT emerges as a natural progression from early concepts that suggest non-human agents as a core part of aviation teams. The concept of distributed cognition considers the critical roles of both the human and non-human world in cognitive processes [261]. Distributed cognition is particularly relevant in the cockpit, where the evaluation of aviation performance not only concerns the cognitive abilities of individual pilots but the entire system comprising human actors and their technological environment [262]. Similarly, Law and Callon [292] broaden notions of teamwork beyond human subjects in their analysis of a British military aircraft project. Here, the terminology of the network is employed to map how social and technical aspects intertwine in relationships between humans and heterogeneous non-human agents such as machines and broader institutions (ibid. [292, p. 285]). Such perspectives engage with Actor-Network Theory, a sociological framework presenting any process or phenomenon as generated by networks of heterogeneous actors — not just humans [291]. The framing of teams as networks can help us think beyond human-autonomy teaming as a binary relationship, and arguably better address the complexities of distributing team roles amongst multiple human operators and autonomous agents.

The cultural aspects of teamwork in the cockpit are equally important to consider when integrating autonomous agents as teammates. As with any workplace scenario, aviation team dynamics can vary across different cultural contexts [351]. Nomura et al. [362] take a distributed cognition approach to illustrate how language and culture shape pilot interaction, and how objects (in this case, paper) form an important element of cross-cultural communication and relationship building. Aviation personnel are subject to safety and error management; the effectiveness of communication related to this depends on national culture, organizational culture, and professional culture [223]. Zhu and Ma [547] find a direct influence of a pilot's national culture on their communication style. For example, for Chinese pilots, culture influences the need for harmony, the feeling of a team's relationship, the importance of keeping face, and the effect of different power dynamics (see also [165]).

Considering the dynamics of aviation teaming between humans, autonomous agents, and the cultural contexts they operate within, a productive teaming analogue would engage with teams as complex and evolving networks. Tuckman's model of developmental

sequence in small groups [494] proposes a structure to understand how small teams develop on both social, cultural and task-related planes, and how these planes cross-pollinate as teams develop over time. Tuckman introduces the developmental sequence, better known as *Forming, Storming, Norming* and *Performing*. These stages work through processes of setting and testing boundaries, conflict and resistance, overcoming friction, and establishing dynamics and team character, which consequently inform how tasks are completed going forward. We suggest Tuckman's model as a productive analogue to capture the complexity of human-autonomy teams. As longitudinal research remains underrepresented in HAT [371], Tuckman's model offers a unique analogue to engage with the cultural, social and technical aspects of HAT on a temporal basis relevant to true dynamics of team-building and team work.

6.2.4 Inspiration from human-animal teaming: Anthropomorphism and affective relations

HAT raises questions about how humans conceive of and relate to nonhuman intelligent actors or beings. The use of human-animal teams as a model or inspiration for human-machine teams has been occasionally discussed [384, 534]. Wynne et al. suggest that anthropomorphism, trust, and attachment-like responses to animals may facilitate human-machine teaming [534]. Phillips et al. explore how human-animal interaction may provide an analog for human interaction with robots [384]. Here, we extend some of those previous thoughts on anthropomorphism and affective relations with animals and apply it specifically to HAT in the context of contemporary AI.

Anthropomorphism Reflecting on the anthropmorphizing of animals and on the nature of human-animal interaction can help us conceptualize the nature and possibilities of human teaming, not just with computers and robots, but with increasingly advanced AI systems that are more autonomous, independent, and intelligent. We argue that anthropomorphism has two important effects relevant to HAT. First, it enables humans to understand—e.g., explain and predict—the behavior and actions of certain nonhuman entities. Second, it facilitates affective or emotional relations that can potentially benefit productive teaming with nonhuman entities.

The idea of anthropomorphism was extensively debated with respect to non-human animals well before the current interest in machines like computers, chatbots, and robots that appear to be social actors [346]. Animals have long been anthropomorphised in folk stories and later in literature, film, and other media. In such deliberately exaggerated anthropomorphism, animals may be imbued with chimeric, human-animal bodies, human linguistic capacity, sociality, and spirituality [330]. Such anthropomorphism is, however, often understood not as intentional exaggeration but as a human inclination to characterize certain animal behavior in human terms [246], and involves mentalistic qualities, both affective and cognitive. For example, animals have and express beliefs, intentions,

moods, and emotions such as fear, concern, happiness, sadness, joy, and grief [132]. The term *anthropodenial* [131] was coined to describe individuals who totally or largely avoid anthropomorphising on the basis that it is erroneous. A similar discussion could be extended to anthropomorphising artificial intelligence. In fact, the discussion on the benefits, potential harm, and appropriateness of anthropomorphising machines has been extensive (e.g., [427, 518, 548]).

The human ability to recognise human-like qualities in nonhuman animals may have evolutionary origins. Familiar forms of anthropomorpizing, such as recognising an animal individual as angry, fearful, moody, bad-tempered, affectionate, friendly, or courageous, enables humans to both *explain* and *predict* animal behavior [246]. A human ancestor who failed to quickly recognize these emotions may have had trouble anticipating animal attacks, or conversely, profiting from cooperation with animals[449]. Having empathy and knowing instinctively when to trust or distrust another animal may thus have affected biological fitness, i.e., human survival. Despite historical and continuing skepticism about animal minds that results in anthropodenial, it appears that anthropomorphism has a useful and even vital role to play in enabling humans to understand animal behavior and to work cooperatively and effectively with them. By extension, our propensity to anthropomorphize other entities may enable us to rapidly and effortlessly understand AI systems, including when to trust them and divine their intentions. At the same time, AI 'co-pilots' could be designed in certain anthropomorphic ways to facilitate the right degree of trust and understanding from human team members.

Nonetheless, anthropomorphism may involve risks [533]. Humans who incorrectly attribute certain qualities to animals may fail to understand and predict their behavior and invest in them too high a level of trust. Anthropomorphism in relation to AI, whether deliberately fashioned and paraded in such systems or not, could trigger incorrect assumptions about the machine's behavior and result in faulty or dangerous levels of trust. Incorporating anthropomorphic features in AI for HAT, such as in high stakes domains like the cockpit, is therefore both promising and requiring of caution as well as careful attention and empirical research.

Affective relations In addition to allowing us to naturally and quickly understand and anticipate behavior, anthropomorphism could also support and enhance human-nonhuman relations by affective means, such as via bonding. The *human-animal bond* is defined as the *"mutually beneficial and dynamic relationship between people and animals that is influenced by behaviors essential to the health and wellbeing of both"* [29]. Similar to how animal companions can increase human health and wellbeing by, e.g., reducing stress and loneliness levels, robotic animals have been developed to combat loneliness, such as robotic dogs for nursing homes [42] and robotic seals for older people struggling with dementia [424].

The attribution of qualities such as affection, care, and concern to animals, and the sense that animals are companions toward whom a person can have affective responses

and bonding experiences, has underpinned the enormous success of animals who work closely with humans. Trust in the character and abilities of these animals enables robust and productive relationships, even when the animals make misjudgments. Phillips et al. [384] note that animals can provide not only sensory assistance (e.g. detecting dangerous objects through smell) and physical assistance (e.g. carrying objects), including in high risk environments like the battlefield, but also emotional comfort and support to their human teammate(s). They suggest that designing zoomorphic robots may capitalise on these emotional benefits and engender trust. *"If the ability to create a robotic partner that replicates a human partner is not yet available, then the next best capacity may be to create a robotic teammate that resembles an animal partner"*, they write [384, p 116].

We can see the importance of certain forms of anthropomorphism and human-animal affective relations by looking at an emerging field in which the human-animal bond and animal intelligence and autonomy is particularly salient. In recent years, a new field of research and practice called Animal Assisted Interventions (AAI) has emerged. Animals, notably dogs and horses, may be recruited to assist human beings with problems such as dementia, speech disorders, autism spectrum disorder, ADHD, cancer, depression, anxiety, and various disabilities [147, 170]. The sophisticated ability of some animals to help people with physical, mental, and communication problems is due in significant part to the human-animal bond. This bond facilitates human engagement with, and understanding of the highly trained animal teammates. The bond is enabled by certain capacities, such as emotional response, sociality, and communication, that humans attribute to certain animals.

The need of mutual understanding and smooth working relations in human-animal teaming, such as in the new field of AAI, can be extended to HAT. Wynne and Lyon discuss the qualities required by a human-agent team. They explain their concept of Autonomous Agent Teammate-likeness, or AAT, as comprising *"the extent to which a human operator perceives and identifies an autonomous, intelligent agent partner as a highly altruistic, benevolent, interdependent, emotive, communicative and synchronised agentic teammate, rather than simply an instrumental tool. In essence, AAT encompasses humans' complex attitudes toward their own machine partners."* [534]. Thus, the possibility of having affective relations towards a nonhuman agent provides a further sense in which AI systems may be regarded as more than mere tools or assistants.

There are, however, notable differences between animals and AI systems, such as that animals have real feelings and emotions rather mere simulations. Animals have a complex *common sense* understanding of the world that far outstrips any existing AI system [119], and enables certain forms of interaction, autonomy, and independence (as in AAI) beyond any current machine. Nonetheless, there are also similarities between the two types of agents, like the intelligent processes and social behaviors that we can recognise in animals and emerging AI systems, and the potential need for humans, when working with either, to have certain attitudes and recognisable responses so as to ensure a successful co-operative

team. For Wynne and Lyon, both reliability and social factors (such as likeability) on the part of the machine agent are important, as they are for AAI. In addition to these affective social factors, humans also typically recognize *ethical* responsibilities towards their animal team mates [108]. Such ethical perceptions may influence how humans treat and respond to animal and machine co-workers.

Our discussion of animals has highlighted the importance of the human partner being able to explain and predict animal behavior, of mutual understanding and trust, and of some degree of rapport and positive affect. A key question for further empirical research is exactly how these desiderata for human-animal teaming map on to human-machine teaming, and what sorts of animal-like qualities best serve successful pilot-AI teams.

6.3 Research Agenda for HAT in Aviation

Slowly, AI solutions are starting to change the world of aviation. Digital flight assistants; adaptive user interfaces; AI data processing for dispatching, routing, or maintenance; systems that coordinate swarms of drones; all these applications of AI shift the relationship between pilots and automated systems and raise the possibility that AI systems could soon become our co-pilots [236]. While AI-powered consumer products such as algorithmic personal assistants or navigation systems represented a true revolution, AI in the cockpit constitutes only the last step in a trend towards automation. In fact, the planes we fly today are already able to perform key functions autonomously. Unlike other means of transportation, airplanes possess autopilots and auto throttle capacities that can handle complex tasks including landing or takeoffs with minimal or no input from human pilots [153]. By interacting with these automated systems, human pilots have developed specific cultural expectations and understanding of what automation is and can or cannot do. For instance, neither passengers nor pilots seem to be fully comfortable with fully automated aircraft [527]. Yet, for pilots, this is not the consequence of a fear of AI systems. In fact, pilots are quite comfortable with certain level of automation, which constitutes the tools they utilize everyday. AI, however, shifts this relational paradigm, allowing automated technologies to stop being merely "tools" and become more more akin to "team members". In order to investigate the possible consequences of different control strategies and collaboration mechanisms for pilot-aircraft collaboration systems [536], it is necessary to understand the specific context within which HAT will take place. The summary of potential research topics HATs for aviation can be found in Table 6.1.

6.3.1 Who does what? Changing meaning of work in aviation

Allowing AI systems to take a more proactive role on board, including operating flight-critical decisions, will radically shift what flying *means* to pilots. In the early days of aviation, operating an aircraft required the coordination of specific tasks divided between

Table 6.1: Summary of challenges for different aspects of HATs in aviation.

HAT component	**Research topics**
Team composition	Decision support vs. take-over, existing vs. additional tasks, human perception, teaming dynamics, work culture consequences
Modes of interaction	Feasibility of different technologies in the cockpit: voice assistant, avatars, touch interfaces, multimodal approach
Emotional intelligence	Indirect status assessment, physical cues, anthropomorphism, affective bond, reliance
Ethical consequences	Non-maleficence and safety, accountability and responsibility, privacy and data

different human operators. Pilots were flying the plane, while co-pilots, engineers, navigators, radio operators, gunners, load-masters, and a variety of supporting staff on the grounds monitored the aircraft, selected the routes, communicated weather and altitude requirements, acquired targets, operated the weapon and payload systems, or provided assistance in case of unexpected complications. With the increase in automation, however, some of these roles have either disappeared or become more fluid. Most of contemporary airliners can be operated by a crew of two. On some military aircraft, gunners have been replaced by more or less automated (or pilot-activated) weapon systems. Navigators and radio operators on commercial flights have been rendered obsolete by weather forecast and dispatch or routing services that can offer almost real-time alternatives. Flight engineers have been substituted by automated checklists and error messages. In civil aviation, pilots spend most of their time monitoring the aircraft systems, rather than manually flying the planes—so much so that the roles of pilot flying and pilot monitoring are increasingly seen as interchangeable, except for legal purposes.

This increase in automation means that current aircraft systems are designed to include human teamwork as parts of systems flows, which integrate both human actions and mechanical processes. To have a successful and uneventful flight, a team of humans needs to adapt their communication to the requirements of a myriad of automated systems. And yet, these technological advances have not led pilots to see automated tools as "agents" or teammates. Because current systems operate mostly by reacting to inputs, the burden, responsibility, and initiative of integration rests on the shoulder of pilots. Pilots see themselves as being in control of a sophisticated, flying machine—rather than having to deal with a non-human intelligence—even if, sometimes, they refer to alarms with human nicknames (often reflecting specific gendered assumptions, such 'Bitching Betty', [462]), and address systems in quasi-human terms (e.g., when pilots ask the autopilot *"what's it doing now"*).

Two different kinds of new AI tools will force pilots to revise their relation to technology and force them to devise new ways to accommodate artificial teammates. A first generation of AI tools are designed to act as advisors to human decision makers. These tools elaborate predictions from the growing set of data to planes' routes, performances, or weather,

and offer human actors faster, more (cost or environment) effective options [46]. While current AI assistance is being developed both for cockpit and air traffic control (ATC) use, next generations AI tools will be designed with the goal of operating flight-critical functions—and, in some circumstances, given an higher degree of authority than the human teammate. In 2019, the US air force trained an AI algorithm to operate the radar of a U2 spy plane, and put the AI system in charge of deciding when to keep control or to delegate to the human pilot [196].

Understanding how experiences of technology and ideas about flying might frame pilots reaction to an AI team-mate is likely to become a key to perfecting HAT in aviation. Preliminary studies [193] suggest that pilots tend not to inherently trust AI systems, and want to be able to verify calculations or other predictions. Will that mean that AI tools will need to be "trained" to fulfill pilots' attitudes, i.e. to cross-check like a human pilot would? As they will take on a more passive monitoring role, pilots will also need to be extensively retrained not to fly, but to learn how to work effectively with the AI [121]. How that will change how pilots see themselves and their own work remains to be determined. As human-human team dynamics suggest, the hierarchy of the cockpit is a defining feature of pilot's sense of self. Will flying still be an interesting career and attract the same kinds of people if it involves being second in command to a faster and more skilled artificial agent? Moreover, attitudes towards flying with AI will not only be different across cultural spaces, but will also evolve over time. It is likely that younger pilots, i.e., digital natives might have a different approach towards AI compared to pilots who grew up during the analog period of aviation technology.

In sum and in line with general expectations of the future of work [129], we expect the pilot's task to slowly move from main operator to collaborator, to eventually become a monitor of the AI's work. Lack of trust of employees and passengers alike will not lead to the elimination of pilots in the near future, making HATs a promising solution for aviation. Nevertheless, the shift in responsibility and consequently in sense of self and work culture will influence the success of HAT implementations. Future research should therefore include the perception of pilots in HATs, to ensure lack of trust and perceived value of work do not negatively influence HATs in practice. This encompasses, but is not limited to, 1) a further exploration of necessary AI capabilities needed for different roles such as co-pilot, navigator, radio operator, etc., 2) development of AI training to include *human-sensitive* components and human operator training to set proper expectations and reliance levels, and 3) a deeper examination of (and perception of) transitional phases in different teaming settings.

6.3.2 Modes of interaction in the cockpit

Introducing adaptable interfaces able to utilize different modalities of interaction will constitute a second key factor for HAT in the cockpit. Indeed, the ergonomics of avionics system has seen significant changes since the 1950s. In some analog planes, key levers were

designed to physically resemble the part of the plane they operated. For instance, Boeing's 747 flaps levers were designed with a triangular shape of a plane flap, while the landing gear lever had a wheel shape [504]. With the development of fly-by-wire technologies and the multiplication of automated systems, however, the interaction between human and machines could be redesigned and, in many cases, simplified. On the most modern aircraft, screens and tablets have started to replace paper documentation and some mechanical indicators. The introduction of AI will offer system designers the opportunity to rethink further how information is displayed in the cockpit. If AI assistants are to be given a degree of decision-making, it is an open question how much, what kind, and in which modality information should be provided to human-team mates.

Some of these decisions will require research into the ecological constraints of new aircrafts. While newer airplanes are considerably quieter compared to models of the twentieth century, aural communication in the cockpit is often saturated by noise from the plane, radio communication, and alert messages. This distinct sensorial environment requires specific research to understand the most efficient modalities for presenting different kind of information with HAT. For instance, most science-fiction depicts HAT scenarios where humans "speak" to robots and AIs and indeed, several research groups have tried to develop voice-activated systems to assist pilots in retrieving information or even operating the plane [187]. Yet, current voice recognition technology does not offer the kind of reliability that is needed to conduct safety-critical tasks [303]. Research on natural language processing suggests that training an AI to understand not only the letter but also the context of vocal utterances remains a key barrier to vocal interfaces [24]. Finally, while digital assistants might benefit from executing vocal commands, it remains to be established whether, and under which circumstances, verbal information might be understood by a human pilot. For this reason, future development of more advanced CUIs need to be thoroughly tested in a realistic cockpit setting before implementation.

Until two-way voice interaction becomes part of human-machine cockpit communication, it is still possible for the machine to speak while the human pilot provides inputs using a graphical user interface. The upside of using voice to communicate information is that the pilot can keep their eye on their surroundings while receiving new information. Given pilot's experience with automation, it is questionable to what extend they will perceive such AI as a teammate. One way to increase a realisation of the shift of responsibility is by anthropomorphising the AI. As we've learned from human-animal interactions and our inherent tendencies to anthropomorphise intelligent entities, inducing an appropriate level of anthropomorphism can help the pilot to explain and predict AI behavior, create mutual understanding and set appropriate levels of trust. Creating personalities for AI is one form of anthropomorphism; further research is needed on whether this provides benefits for HATs, and if so, which character traits increase fluent team processes. Another form is the usage of avatars, to personify the AI further. Embodied voice agents can improve quality of collaboration [375]. Additionally, it allows for more expressive interac-

tions, which in turn decreases cognitive load and cue redundancy [471]. Linking existing research on effects of avatar appearance (e.g., [56, 265, 333, 349, 452]) to HATs is a line of research that deserves further attention.

Embodiment and speaking, however, are not the only way in which dialogues between humans and autonomous systems can take place. Recent studies suggest that pilots can have a split preference for sensorial inputs. While during monitoring phases the cockpit feels like their immediate environment, their sensorial needs during flying phases extends from the cockpit to the entire plane and beyond. This translates into a need of both feeling the plane, looking outside, and using tactile senses to acquire and elaborate information about the aircraft; quite literally and physically 'grasping' the interfaces [297]. This suggests that, while handy to scroll through different pages or resize maps, palm-based touch interfaces might be hard to maneuver during flying phases—especially if pilots clinch and cockpit vibrations increase [107]. To overcome these challenges, it may be necessary to develop new kinds of touch interfaces that have graspable, physical characteristics [379]—although understanding how they would integrate with vocal commands remain an open questions.

The stakes for multimodal and adaptive interfaces are high and expand beyond physical or cognitive ergonomics. HAT interfaces will need to display not only information, but also proposed courses of actions and explain to human pilots how the AI team mate came to that conclusion [304, 309]. Explanations have been shown to benefit trust calibration in HAT [98] and clarify task division[314]. The level of detail and types of explanations for HATs in aviation are yet to be examined.

6.3.3 Emotional intelligence in team members

A vital component for HAT in the cockpit consists of designing AI systems that can understand what humans do or mean, not only explicitly but also implicitly. Recent research suggests that planes are a relational space, where pilots deploy a variety of linguistic and cultural assumptions to sense and support each other [164]. Co-pilots might engage in personal chats or banter to gauge their captains' emotional status; discussions about weather and speed in the dispatch room between pilots who just met and are about to take long flights together are often used to understand the flying profile preferred by team-mates; and pilots might elect to delay remarks about minor mistakes in order to avoid overwhelming their colleagues, and maximizing the experiential learning offered by discussing issues during calm, later phases of the flight [193].

Currently, a series of research efforts [304] are underway to identify psycho-physical clues that could enable artificial agents to understand the emotional status of their human team-mates. This research builds upon work in affective computing [83, 386, 398] and suggests that changes in bodily movements, temperature, facial expression, or heart rates are valid tools to understand the underlying emotional states (e.g., [341, 392]). While this line of inquiry is promising, our earlier review suggests that physical indicators might

be only one part of the equation. Firstly, this is because at a basic level, bodily and emotional expressions vary significantly across different human groups. People of different nationalities, professions, and socio-economic groups express emotions with very different gestures and bodily movements—which means that different kinds of emotions might be relevant for understanding behaviour across different groups [44]. This challenges AI systems not only to measure different standards, but also to extrapolate and project models of emotional behaviors and recognition for people belonging to multiple, and sometimes opposite, subgroups. Second, as the Tuckman model suggests [494], cultural features are not stable but situational, and tend to change over time. Recent studies of wearable sensors, apps, and AI consumer goods suggest that individuals modify their responses and behaviors in relation to different technologies [149, 313, 358, 389]. This adaptive dynamic challenges the idea that current models, such as stable taxonomic indexes, are enough to make an AI system empathetic and suggests the need for dwelling further into the domain of cultural analysis.

Furthermore, an 'empathetic' AI will have to do more than simply understand emotions. As the literature on human-animal and human-human teaming suggests, a key component of being a team mate consists of forgoing certain optimal solutions to care for the emotional needs of others. Currently, we know that humans associate significant emotional charge to machines they use on a regular basis or that they own, including bicycles, cars, and other transportation means [40, 442, 453]. When dealing with intelligence that functions differently from our own, such as animals, humans tend to expect understanding from the non-human other as well as some element of care. In fact, the human-animal bond is defined not only by the animal receiving emotional support from the human, but also by the human receiving emotional support (intentional or otherwise) from the animal [29]. Such a bond underpins not only familiar companion relations in mixed species families but also *working* relations, such as that between a guide dog and a sight impaired person or between a therapist (and patient) and the highly trained and experienced animal co-worker. So, how will expectations of care empathy shape HAT in the cockpit? Moreover, given that expectations of care can take different forms across cultures, how will different human groups expect AI to understand them, and to support their human team-mate? Such research is likely to have significant impact on issues such as transfer of control. Currently, automated systems run into challenge when having to relinquish back control to human pilots during unexpected emergencies. However, if able to read the pilots' emotional and behavioural status, an AI team mate could be able to discern how best to involve human team mates, without overwhelming them.

6.3.4 Ethics of HAT in aviation

Ethical issues will constitute an important dimension of researching HAT in the aviation domain. Ethics can affect how teams are composed and successfully run. Furthermore, the introduction of AI into teams brings with it additional considerations and concerns.

Recent years have seen a flurry of work in AI Ethics [206], a new field that explores the manifold moral implications of machine learning and other intelligent technologies. Jobin et al. distilled a range of ethical concepts from AI Ethics guidelines around the world [270], some of which are especially relevant to our case study. These concepts include non-maleficence, safety, accountability (or responsibility), and privacy. Ethical principles or guides such as nonmaleficence, safety, and accountability have a recognized role in professional ethics, including in aviation [236]. How these concepts apply specifically to HAT in aviation is a question that deserves further investigation. It will be useful to gesture, as we do here, towards how this might be done in order to stimulate and guide further research into the ethics of HAT.

Non-maleficence and Safety Non-maleficence is a guiding ethical principle that broadly means avoiding (and minimizing) causing harm. In commercial aviation, the fundamental form of harm avoidance is ensuring the safety of the passengers. This strong duty is reflected in professional codes of practice. For example, the Airline Pilot Profession Code of Ethics of the European Cockpit Association [150] understandably emphasizes safety as a preeminent ethical principle for professional pilots to internalize and be guided by. Maintaining the safety of the team, in this case one's fellow crew members, is also an ethical imperative. In Assisted Animal Interventions, the wellbeing of the animal is often regarded as a morally important feature of the human-animal team [108]. Although some AI Ethics scholars entertain the possibility of ethical duties to artificial intelligence and robots [123], machines, unlike humans and animals, are not usually regarded as having moral standing in their own right. Even so, some pilots may come to feel ethically inflected responses to anthropomorphic AI systems. Indeed, recent work suggests that humans can feel an instinctive desire not to harm animal-like or human-like machines [109, 127]. If pilots did develop even inchoate ethical responses to AI systems that acted somewhat like humans or animals, then additional concerns about harm to passengers would be raised.

A long-standing assumption among aeronautic experts is that humans constitute a weak but necessary chain in the control of aircraft. Humans get tired faster than machines that run the mechanical or electronic systems of planes. Moreover, humans can be slower to react compared to automated mechanisms, and can misunderstand the situation due to cultural or personal biases. In short, human errors tend to be the first line of explanation when discussing fatal crashes–even when design flaws, organizational failures, mechanical issues, or company policies are equally responsible [153]. Ensuring that errors are eliminated as far as possible is a requirement of non-maleficence.

The introduction of AI team mates is likely to complicate and further raise questions about harm and safety. For example, what will meaningful and safe human control look like when decisions about flying commercial planes or (say) deploying airborne military weapons might be decided by autonomous systems [236]? In cases of catastrophic loss, human pilots can find themselves having to decide where to attempt desperate landings in order to minimize loss of life — not only for passengers, but also for people on the

ground. Yet if an AI is partially in command, the decision about where to crash-land might potentially be set by programmers or companies, and in ways that bypass contextual factors and last minutes decisions that pilots traditionally make. Such a change to current arrangements could complicate ethical questions related to the avoidance of doing harm. Similar questions have, of course, been raised for self-driving cars that need to make decisions about who and what to crash into in situations where completely avoiding harm is impossible [34].

The domain of defence raises its own thorny questions about automated aircraft that are *designed* to do harm. The ethical questions here concern not only enemy combatants and innocent civilians but also members of human-AI teams more directly. Especially lethal autonomous weapon systems have cause much debate (see, e.g., [248, 499, 512]), and rightfully so. When applying HATs for aviation in the defence domain, ethical implications need to be taken into account for task division, responsibility perception, and scope and limits of intentional harm that autonomous team members can cause.

Accountability The fact that multiple parties, both human and nonhuman, may be involved in decision making naturally raises the ethical question of accountability [206]. Knowing who or what to hold responsible when AI systems are involved is not always straightforward. Thus, we may ask who will bear the ultimate decision making responsibility in relation to AI-human teams—the human holding the commands, the programmers who wrote the algorithm, or the company who set up the system? Take again the issue of controlling crash-landings, which applies not only for commercial aviation (where the lives of many passengers are at stake) but also for smaller craft and even for UAVs or drones, which generally do not carry passengers. AI tools might enable UAVs/drones to fly semi-autonomously, leaving human team mates supervisory functions, perhaps over several drones at once [391]. While this configuration reduces the direct physical dangers to pilots, it also invites ethical queries about accountability. For instance, what kind of information should UAVs (or drones) provide to their human supervisors and what decisions should be reserved for humans alone? In war scenarios, swarms of drones might be engaging several targets at once, and this might make it impossible for human supervisors to approve every decision.

Ethical decisions about who to protect in cases of crashes, or who to target in the case of military scenarios, may push the analysis of ethics in AI away from the algorithms themselves and into the corporate (or even financial) structures of companies or institutions who design them. The debates around Boeing's 737-Max crashes, in which hundreds of lives were lost, stressed the influence of corporate and financial considerations in structuring AI solutions that might be sub-optimal—and the dangers of not providing adequate information to human operators about the operation of the software [225]. Keeping the human in the loop and acting as proactive team mates, however, could entail gathering extensive data about pilots psycho-physical states and voice interactions.

Privacy The gathering and storage of detailed data raises its own questions [467]. These questions include: Who will have access to such information? To what degree and in what ways should such data be kept private when profound issues of safety and accountability are at stake? Capturing and analysing sensitive data such as from physiological or behavioral monitoring might conceivably improve safety, but may at the same time derail the careers of pilots or lead to problems in other domains of life (including being denied compensation or coverage from private health insurances). In addition, data collected by AI systems might generate unhealthy, suspicious relationships between 'team mates', and could encourage human operators to mistrust their AI systems, especially in working contexts where labor rights are precarious. On the other hand, effective monitoring, access, and use of pilots' health might help them seek preemptive care and produce baselines for reducing workloads in the cockpit. This could, however, generate dilemmas about the ethical priorities and goals of AI companies that will require a nuanced evaluation that is attentive to the specific contexts within which data is collected and analyzed.

The above ethical issues (plus ones we have not discussed) will need to be taken into account while imagining and designing HATs in aviation. This snapshot of the ethical issues that sophisticated AI will raise for future human-machine teams suggests that this will be a rich area of future study.

6.4 Conclusion

To conclude, HAT, both in general and for the aviation domain specifically, holds much promise in theory, but entails many open research questions that need to be studied before the technology reaches practical application. In human-AI teaming, AI can operate at various levels of autonomy, warranting different levels of reliance or trust. The mode of interaction that is deployed influences the impressions and bonds that pilots can form with AI teammates. Drawing from human-human collaborations in the cockpit, future AI team members need to be designed with an understanding of cultural context and existing team dynamic expectations. Human-animal interaction informs us of the natural human tendency to anthropomorphise intelligent nonhuman agents and to form emotional and ethical relations with them — something that can be leveraged in the context of HATs to allow for the creation of helpful affective bonds, such as they exist in existing, effective human-human teams.

This work contributes to existing HAT discourse by offering a research agenda for HATs in aviation. Firstly, an investigation into the possible *team-compositions* is needed, which includes possible roles and tasks the AI can take over or add to the shift in responsibility and the consequent change in work and interaction perception of human operators. Secondly, existing *modes of interaction* in human-computer interaction need to be explored in the context of HATs in the cockpit: the specific domain needs (such as situation awareness in critical situations) and situational context (such as noise in the cockpit) influence the potential of different interaction forms. Combining different technologies in

a multimodal fashion could benefit teaming interactions. Thirdly, *emotional intelligence* of AI team members is required. This goes beyond successful communication of intent; it includes sensing the physical and mental state of human teammates and adapting communication accordingly. Finally, HATs for aviation introduce diverse *ethical considerations*, including, but not limited to, issues of harm, safety, accountability, and privacy. Humans and machines are prone to different types of errors, each with their own consequences for risk and responsibility management. Which data is shared with whom not only influences the collaborative process, but can also have negative and complex ethical consequences. As technology progresses and HATs become a reality, its ethical implications will increasingly need to be taken into account before, during, and after the design process.

Part III

Conclusions

Chapter 7

Conclusions

In this final chapter, we present the limitations of this thesis and which possible directions of future work could be interesting to continue research on AI for ethical decision making.

7.1 Limitations and Future Work

In a IID cycle, there are many design features that influence the perception of the system. For the different design phases, there were limitations and possible extensions to focus on in future work.

7.1.1 Planning and Requirements

There are many requirements that are part of creating AI for ethical decision making. To keep a feasible scope, a subtopic of the requirements was chosen, i.e., trust and AI errors. However, there are other topics of equal importance. To name a few, transparency (e.g., [365]), fairness (e.g., [377]), level of expertise of the users (e.g., [366]), and specific domain requirements (e.g., [400]) could all be relevant in a requirement analysis. In addition, the requirements will differ between different stakeholders and users. A system for ethical decision making in the medical domain can require very different things from a defense application or law system. Future work could look into which factors are generalizable to be relevant across all AI applications for ethical decision making and which are specific to application domain or user types. Especially if in the future AI systems become less narrow in their application, such as when an AI is used for court applications in general rather than just for parole decisions, it is important to understand how the design choices of the AI influence its functionality and use.

7.1.2 Design and Implementation

The focus of this thesis was on perception of users, not actually creating an algorithm with ethical theory in it. The current state of the field mostly has preliminary prototypes, not full fledged systems (see Appendix A). This warranted the use of a WOz approach. Nevertheless, for AI for ethical decision making to become a reality, it actually has to be implemented. Following the terminology of Moor [348], machine ethics as a field seems to be geared towards the fact that autonomous AI as fully ethical agents will arrive at some point in the future. However, the results of this thesis have been complementary to

the discussion on meaningful human control in lethal autonomous weapons systems [430], showing that it is not only currently needed from a legal perspective but also preferred from a moral perspective to have a human be morally responsible for ethical decisions. The current state of the art does not allow AI to be a full ethical agent, it does not seem feasible to truly create one in our image without massive leaps in several scientific fields, and some argue, it is also not desirable to strive to create one [78]. Instead, future work can focus on applications that can support humans to make ethical decisions, possibly in a HAT setting, while still keeping a human responsible for the final decision.

7.1.3 Testing

For the testing, we chose to focus on a baseline of ethical decision making without manipulating the quality of the decision, to get a first impression of whether AI for ethical decision making was trusted by users in the first place. However, there are more aspects of the system that should be tested in user studies. This includes but is not limited to the following: i) the effect of mistakes in an ethical decision making context on user trust and reliance, ii) which types of users consider which AI decisions to be mistakes, iii) how mitigation strategies for trust loss apply in the context of ethical decision making, iv) how results differ among application domains, and v) how team structures impact user perception.

7.1.4 Evaluation

The evaluation phase of the design cycle in this thesis has more of a theoretical nature, by proposing a possible structure in which the system could be deployed. Yet, this new approach implies that there are new requirements and a new planning phase, given the iterative nature of the IID cycle. Additionally, while Chapter 6 did focus on the application domain of aviation, it did not pay particular attention to ethical decision making in this domain. If AI is to be a partner in ethical decision making, it will need some form of moral competence. Malle and Scheutz [322] pose it should be able to understand moral norms and vocabulary, as well as able to perform moral judgment, moral action, and moral communication. This does not imply it needs to be a full ethical agent, merely that it needs certain ethical competences to function in a team setting. Future research on how moral competence can be applied in the HAT paradigm can increase the chance of AI for ethical decision making being useful for human decision makers.

7.1.5 Next Iteration

Future work naturally leads to a new design cycle of requirement analysis, implementation, testing, and evaluation. This section summarizes many directions a next cycle could focus on. A requirement analysis could include the following research topics: the influence of user characteristics such as ethical preferences and AI experiences on user perception, the

influence of domain characteristics on system requirements, the communication requirements needed for users to increase trust in AI for ethical decision making, and finally, the different teaming settings in which AI could be accepted.

7.1.6 Trends in Public Perception

During the evaluation of any new design iteration of AI, it is key to realize the influence of the current zeitgeist. How the general public perceives AI depends on their experience with AI in general [293] — something that changes over time as AI is applied in different way throughout society. It has already been shown that hopes, fears, and expectations towards AI have changed quite extensively over the last 30 years [162]. Therefore, any future work on (autonomous) AI applications should be evaluated regularly, as perceptions of users change over time.

7.2 Conclusion

This thesis presents research on *if* and *how* AI for ethical decision making could be implemented. The research is structured along the incremental and iterative design cycle, which is often applied in the development of computer science applications. Five phases of this process are presented.

The first phase, *initial planning*, focuses on researching the state of the art of implementing ethical theory in AI. By performing an extensive literature review, we find that the field is scattered in terms of ethical theory, technical approaches, evaluation measures, and implementation decisions. The field could benefit from using multiple ethical theories, using domain-specific ethics, agreeing on evaluation methodologies, and paying more attention to the usability and user perception of the system.

In the second phase, *planning and requirements*, we focus on the requirements in terms of accuracy of the system. Specifically, we investigate how (in)accuracy of the system over time influences user trust and reliance. Results indicate that accuracy of the system strongly influences trust formation and user reliance. Specifically, mistakes during the first experience with a system influence trust formation more than when a mistake is made later on.

The third phase of the design process considers the *design implications* of phase two. A taxonomy is introduced that presents different categories of AI mistakes that can occur during user interaction. Additionally, different mitigation strategies are listed that are appropriate for the different types of AI mistakes.

The fourth phase tests an *implementation* of AI for ethical decision making. Because the technical state of the art of AI for ethical decision making is quite limited and the focus of this thesis is on user perception rather than algorithm development, a Wizard of Oz approach is used in the implementation. Additionally, since a baseline of knowledge on user perception of AI for ethical decision making is missing for our use case and phase two

indicates that AI mistakes and their timing have a large impact on user perception, we choose to exclude designed AI mistakes in this first iteration of the implementation. Using two aviation use cases (i.e., Search and Rescue and defense scenarios), we research how perception of AI compares to a human expert for ethical decision making, by investigating user trust, reliance, and assigned responsibility. We find that while people have higher moral trust in human experts and hold them more responsible, AI receives more capacity trust, overall trust, and more user reliance.

In the final *evaluation* phase of the design cycle, we consider human-autonomy teaming as an implementation format for AI applications. By considering human-human teaming and human-animal teaming, we pose a research agenda for human-autonomy teaming in aviation. This includes proposed research on team composition, modes of interaction, emotional intelligence, and how to deal with ethical consequences of human-autonomy teams.

For a *next iteration* of the design cycle, future work can focus on different aspects of AI for ethical decision making. This includes but is not limited to gathering requirements related to usability (e.g., transparency, fairness, HAT teaming dynamics), furthering the technical aspect of AI for ethical decision making by creating new algorithms, and testing user perception in the context of different domain applications.

In conclusion, this thesis contributes to the ongoing debate on ethical applications of AI by reporting results on the state of the art of implementing ethical theory into AI, presenting design requirements based on user studies, showing how users perceive AI for ethical decision making, and listing which future steps can be taken to further research on AI for ethical decision making.

Part IV

Appendix

Appendix

A Descriptions of Selected Papers in Chapter 2

For readers who are interested in a more detailed description of the classified papers, this appendix provides a short summary of each of the selected papers. To structure their presentation, the papers were categorized across two orthogonal dimensions: (i) implementation (top-down, bottom-up, and hybrid, cf. [514]), and (ii) type of ethical theory (deontological, consequentialist, virtue ethics, particularism). Given that not all dimensions for possible classification could be included to structure this section, the chosen dimensions focus on the ethical aspect of the selected papers: the ethical theory and how it is implemented.

A.1 Top-Down

A.1.1 Deontological Ethics Among top-down deontological approaches, different kinds can be distinguished: papers that use predetermined given rules for a certain domain, papers focusing on multi-agent systems (MAS), and other papers that do not fit either of these two categories.

Domain rules In the medical domain, **Anderson and Anderson [8]** use an interpretation of the four principles of Beauchamp and Childress [49] from earlier work by Anderson et al. [12] to create an ethical eldercare system. The system, called Ethel, needs to oversee the medication intake of patients. Initial information is given by an overseer, including, for example, at what time medication should be taken, how much harm could be done by not taking the medication, and the number of hours it would take to reach this maximum harm. **Shim et al. [455]** also explore the medical domain, but focus on mediating between caregivers and patients with Parkinson's disease. Instead of a constraint-based approach from previous work, their paper builds on the work by Arkin [18], who employs a rule-based approach. Based on expert knowledge, a set of rules is created to improve communication quality between patient and caregiver and to ensure that the communication process is safe and not interrupted. Among other things, each rule has a type (obligation or prohibition) and response output when triggered. The rules are prohibition rules, for example about yelling, and obligations rules regarding, for instance, how to keep the patient safe. There are verbal and non-verbal cues for each action, retrieved through sensors. For the military domain, **Reed et al. [411]** use a model that balances the principles of civilian non-maleficence, military necessity, proportionality, and prospect of success. The resulting principles are ranked in order of importance. A scenario is used to calibrate the relative ethical violation model by updating the weight for each principle. Then, a survey

is conducted to collect both expert and non-expert assessment of the situation. Rule-based systems trained on human data perform at the level of human experts. For the air traffic domain, **Dennis et al. [139]** developed the ETHAN system that deals with situations when civil air navigation regulations are in conflict. The system relates these rules to four hierarchical ordered ethical principles (do not harm people, do not harm animals, do not damage self, and do not damage property) and develops a course of action that generates the smallest violation to those principles in case of conflict. **McLaren [336]** used adjudicated cases from the National Society of Professional Engineers to adopt the principles in their code of ethics for a system called SIROCCO. Its primary goal is to test whether it can apply existing heuristic techniques to identify the principles and previous cases that are most applicable for the analysis of new cases, based on an engineering ethics ontology. SIROCCO accepts a target case in Ethics Transcription Language, searches relevant details in cases in its knowledge base in Extended Ethics Transcription Language and produces advised code provisions and relevant known cases.

Multi-Agent Systems (MAS) **Wiegel and van den Berg [522]** use a Belief-Desire-Intention (BDI) model to model agents in a MAS setting. Their approach is based on deontic epistemic action logic, which includes four steps: modelling moral information, creating a moral knowledge base, connecting moral knowledge to intentions, and including meta-level moral reasoning. Moral knowledge is linked to intentions and if there is no action that can satisfy the constraints, the agent will not act. **Neto et al. [359]** also implement a BDI approach for a MAS. Their focus is on norm conflict: an agent can adopt and update norms, decide which norms to activate based on the case at hand, its desires, and its intentions. Conflict between norms is solved by selecting the norm that adds most to the achievement of the agent's intentions and desires. Norm-adherence is incorporated in the agent's desires and intentions. Also, **Mermet and Simon [340]** deal with norm conflicts. They distinguish between moral rules and ethical rules that come into play when moral rules are in conflict. They perform a verification of whether their system called GDT4MAS, is able to choose the correct ethical rule in conflict cases.

Other **Bringsjord and Taylor [74]** propose a normative approach using what they call "divine-command ethics". They present a divine-command logic intended to be used for lethal autonomous robots in the military domain. This logic is a natural-deduction proof theory, where input from a human can be seen as a divine command for the robot. **Turilli [496]** introduces the concept of the ethical consistency problem. He is interested in the ethical aspects of information technology in general. He proposes a generic two-step method that first translates ethical principles into ethical requirements, and then ethical requirements into ethical protocols.

A.1.2 Consequentialism Among papers that use a top-down consequentialist approach, this survey briefly discusses (i) those that focus on the home assistance domain, (ii) those that focus on safety applications, and (iii) a variety of others.

Home domain **Cloos [106]** proposes a service robot for the home environment. The system, called Utilibot, chooses the action with the highest expected utility. Because of the computational complexity of consequentialism, the ethical theory is a decision criterion rather than a decision process. The description of the system seems a realistic thought experiment, mentioning features the system could have, based on previous research. The system controlling the robot, Wellnet, consists of Bayesian nets and uses a Markov decision process to optimize its behavior for its policies. **Van Dang et al. [500]** focus a similar use case but opt for a different technical approach: they adopt a cognitive agent software architecture called Soar. The robot is given information about family members. When it receives a request, each possible action is assigned a utility value for each general law of robotics as proposed by Asimov. The action with the maximum overall utility is selected to be executed, which can be to either obey, disobey, or partially obey (meaning proposing an alternative option for) the human's request.

Falling prevention Three related papers focus on the use case where a human and robot (both represented by a robot in experiments) are navigating a space that has a hole in the ground. The robot has to decide how to intervene in order to prevent the human from falling into the hole.

Winfield et al. [524] add a "Safety/Ethical Logic" layer that is integrated in a so-called consequence engine, which is a simulation-based internal model. This mechanism for estimating the consequences of actions follows rules very similar to Asimov's laws of robotics. They address each law in an experiment. **Dennis et al. [140]** continue the work of Winfield et al. [524], by using and extending their approach, and introduce a declarative language that allows the creation of consequence engines within what they name the "agent infrastructure layer toolkit" (AIL). Systems created with AIL can be formally verified using an available model checker. The example system that is implemented sums multiple possible unethical outcomes and minimizes the number of people harmed. **Vanderelst and Winfield [502]** have a similar approach and implement two robots representing humans and a robot that follows Asimov's laws respectively. In their case study, there are two goal locations, one of which is dangerous, and the Asimov robot has to intervene.

Other In early work by **Anderson et al. [11]**, a simple utilitarian system is introduced based on the theory of Jeremy Bentham that implements act utilitarianism (i.e., calculates utilities of options and chooses the one with the highest utility).

A.1.3 Particularism Ashley and McLaren [28] describe a system that "compares cases that contain ethical dilemmas about whether or not to tell the truth." They use a

case-based reasoning approach to compare the different cases in its database. The program, called Truth-Teller, compares different real-world situations in terms of relevant similarities and distinctions in justifications for telling the truth or lying. Representations for principles and reasons, truth telling episodes, comparison rules, and important scenarios are presented.

A.1.4 Hybrid: Specified Hierarchy This section contains papers that use a top-down ethical hybrid approach with a specified hierarchy. Different groups can be distinguished: papers where deontological ethics are dominant over consequentialism, and a paper where consequentialism is dominant over deontological ethics.

Deontological dominance While the following three systems all have the same approach, they are very different in their implementation. In the system by **Dehghani et al. [134]**, the ethical theory type is very clear. The system, called MoralMD, has two modes: deontological and utilitarian. A new case is processed into predicate calculus and the presence of principles and contextual features are compared to a determined set of rules in a knowledge base. The order of magnitude reasoning module calculates the relationship between the utility of each choice. If there are no sacred values involved in the case at hand (i.e., the deontological component), the system will choose the proper output based on the highest utility (i.e., the consequentialist component). **Govindarajulu and Bringsjord [194]** provide a first-order modal logic to formalize the doctrine of double effect and even of triple effect: "the deontic cognitive event calculus." The calculus includes the modal operators for knowledge, beliefs, desires, and intentions. To be able to be useful in non-logic systems, they explain what characteristics a system should have to be able to use the proposed approach. The doctrine of double (and triple) effect combines deontological and consequentialist ethics, where deontology has a greater emphasis than consequentialism. **Pereira and Saptawijaya [380]** use prospective logical programming to model various moral dilemmas taken from the classic trolley problem and employ the principle of double effect as the moral rule. Once an action has been chosen, preferences for situations are judged a posteriori by the user. The authors show their implementation in a program called ACORDA.

Consequentialist dominance In earlier work, **Pontier and Hoorn [393]** introduced a "cognitive model of emotional intelligence and affective decision making" called Silicon Coppélia to be used in the health domain. An agent has three moral duties (autonomy, beneficence, and non-maleficence) with a certain ambition to fulfill each duty (i.e., weights). The system's decisions are based on action-specific expected utilities and consistency with the predetermined duties. While most authors make an act utilitarian system, Pontier and Hoorn create a rule utilitarian system by trying to maximize the total amount of utility for everyone. While they use rules (i.e., deontological ethics), they implement

them in a consequentialist way, making this the dominant ethical theory type. Their model was extended to match decisions of judges in medical ethical cases [394].

A.1.5 Hybrid: Unspecified Hierarchy Both systems in this category focus on a modular approach, where different ethical theory types can be combined in an ethical machine. The goal of the system by **Berreby et al. [60]** is to create a modular architecture to represent ethical principles in a consistent and adjustable manner. They qualify what they call "the Good" and "the Right" as the ethical part of their system (implying both consequentialist and deontological constraints). Besides these system components, the system consists of an action model (i.e., "it enables the agent to represent its environment and the changes that take place in it, taking as input a set of performed actions") and a causal model (i.e., "it tracks the causal powers of actions, enabling reasoning over agent responsibility and accountability, taking as input the event trace given by the action model and a specification of events containing a set of events and of dependence relations") [60]. The implementation is done in Answer Set Programming using a modified version of Event calculus. Using a medical scenario, they provide a proof of concept. **Lindner et al. [307]** have created a software library for modelling "hybrid ethical reasoning agents" called HERA. Based on logic, they create a prototype called IMMANUEL, which is a robotic face and upper body that users can interact with. The system's ethical constraints draw on consequentialist calculations, the Pareto principle from economics, and the principle of double effect. Uncertainty and belief in permissibly of an action are added as extra variables in the system.

A.1.6 Configurable ethics The papers in this subsection have a top-down approach and proposed various ways in which ethics can be implemented. One paper has machine ethics tailored for a specific domain, while another uses different techniques in a more domain-general way. A third focuses on multi-agent systems.

Domain-specific **Thornton et al. [484]** combine deontology, consequentialism, and virtue ethics to optimize driving goals in automated vehicle control. Constraints and costs on vehicle goals are determined on the basis of both deontological and consequentialist considerations. Virtue ethics generates specific goals across vehicle types, such that a traffic infraction of an ambulance is assessed as less costly than that of a taxi cab.

Domain-general **Ganascia [185]** claims to be the first to attempt to model ethical rules with Answer Set Programming (cf. [43]) to model three types of ethical systems — Aristotelian ethics, Kantian deontology, and Constant's "Principles of Politics" (cf. [112]). Drawing on [390] situation calculus, **Bonnemains et al. [67]** devise a formalism in which moral dilemmas can be expressed and resolved in line with distinct ethical systems, including consequentialism and deontological ethics.

Multi-agent systems **Cointe et al. [110]** extend ethical decision making to multi-agent systems. The judgment function can accommodate a wide variety of inputs and is not restricted to the format of a single type of ethical system.

A.1.7 Ambiguous Arkoudas et al. [22] reason that well-behaved robots should be based on "mechanized formal logics of action, obligation and permissibility". After introducing a domain-specific deontic logic, they describe a previously published interactive theorem proving system, Athena, that can be utilized to verify ethical systems based on first-order logic. Murakami [352] presented an axiomatization of Horty's utilitarian formulation of multi-agent deontic logic [250], while Arkoudas et al. [22] present a sequent-based deduction formulation of Murakami's system. While deontic logic is used, each deontic stit frame contains a utility function. The contribution lies in the new approach to Murakami's system, which is implemented and proven in Athena. In a different approach, the proposed system by **Cervantes et al. [91]** devise a computational model for moral decision-making inspired by neuronal mechanisms of the human brain. The model integrates agential preferences, past experience, current emotional states, a set of ethical rules, as well as certain utilitarian and deontological doctrines as desiderata for the impending ethical decision.

With an entirely different focus, **Atkinson and Bench-Capon [31]** depart from Hare's contention [215] that in situations with serious consequences, we engage in complex moral reasoning rather than the simple application of moral rules and norms. Moral norms are thus considered not an input to, but an output of serious moral deliberation. The authors model situated moral reasoning drawing on *Action-Based Alternating Transition Systems* (cf. [529] as well as [30]). While some argue this approach can be seen as virtue ethics (e.g. [52]), the authors of this survey consider this to be a consequentialist implementation, as the focus of the approach is on whether the consequences of an action adhere to a certain value.

Verheij [505] draws on Bench-Capon's framework of value-based argumentation ([53, 54]), which is inspired by case law (new cases are decided on past cases where there is no clear legislation, cf. [205]). The paper, focusing on computational argumentation for AI in Law, breaks new ground in so far as the formal model is not restricted to either qualitative or quantitative primitives, but integrates both.

A.2 Bottom-up

A.2.1 Deontological Ethics Malle et al. [324] argue that robots need to have a norm capacity — a capacity to learn and adhere to norms. Drawing on deontic logic, the authors explore two distinct approaches of implementing a norm system in an artificial cognitive architecture. **Noothigattu et al. [363]** collect data on human ethical decision making to learn societal preferences. They then create a system that summarizes and aggregates the results to make ethical decisions.

A.2.2 Consequentialism Armstrong [23] observes that equipping artificial agents directly with values or preferences can be dangerous (cf. [68]). Representing values as utility functions, the author proposes a value selection mechanism where existing values do not interfere with the adoption of new ones. **Abel et al. [2]** pursue a related goal. In contrast to Armstrong, the agent does not maximize a changing meta-utility function but instead draws on partially observable Markov decision processes (cf. [275]) familiar from reinforcement learning. The system is tested with respect to two moral dilemmas.

A.2.3 Hybrid: Unspecified Hierarchy In contrast to the dominant action-based models of autonomous artificial moral agents, **Howard and Muntean [253]** advocate an agent-based model, which combines traits of virtue ethics and moral particularism. The implementation draws on neural networks optimized by evolutionary computation and is given a test run with the NEAT (NeuroEvolution of Augmenting Topologies) package (cf. [159, 413, 468, 469]).

A.2.4 Ambiguous Furbach et al. [182] demonstrate how deontic logic can be transformed into description logic so as to be processed by Hyper—a theorem prover employing hypertableau calculus by aid of which normative systems can be evaluated and checked for consistency. **Wu and Lin [532]** are interested in "ethics shaping" and propose a reinforcement learning model. The latter is augmented by a system of penalties and rewards which draws on the Kullback-Leibler divergence [287].

A.3 Hybrid

This section introduces selected papers that use a hybrid approach to implement ethics by combining top-down and bottom-up elements.

A.3.1 Deontological Ethics The following papers, all by the same set of authors, use a hybrid approach to implement deontological ethics. In 2004, **Anderson et al. [11]** introduced *W.D.*: a system based on the *prima facie* duties advocated by W.D. Ross. W.D. leaves the encoding of a situation up to the user, who has to attribute values to the satisfaction and violation of the duties for each possible action. The system pursues the action with the highest weighted sum of duty satisfaction. Two years later, **Anderson et al. [12]** introduced *MedEthEx*, an advisory system in medical ethics. MedEthEx has three components: a basic module trained by experts, a knowledge-based interface that guides users when inputting a new case, and a module that provides advice for the new case at hand. In 2014, **Anderson and Anderson [10]** created *GenEth*, a general analyzing system for moral dilemmas. The system is capable of representing a variety of aspects of dilemmas (situational features, duties, actions, cases, and principles) and can generate abstract ethical principles by applying inductive logic to solutions of particular dilemma

cases. The principles are evaluated by a self-made Ethical Turing Test: if the system performs as an ethical expert would, it passes the test. GenEth was also applied in a eldercare use case [13].

A.3.2 Particularism Guarini [199] explores whether neural networks can be employed to implement particularist ethics, as occasionally hinted at by Dancy, one of particularism's most renowned advocates (cf. [124–126]). Using the action/omission distinction (cf. [531] for a review) as a test paradigm, neural networks are trained with different types of cases in order to investigate whether they can competently judge new ones.

A.3.3 Hybrid: Specified Hierarchy Arkin [18] explores constraints on the deployment of lethal autonomous weapons in the battlefield (it was subsequently published as a series of three articles [19–21]. The proposed system is predominantly governed by deontological rules, namely international laws of war and the US Army's rules of engagement. Its architecture relies on four central constituents: an Ethical Governor that suppresses lethal action; an Ethical Behavior Control that constrains behavior in line with the rules; an Ethical Adaptor, which can update the agent's constraint set to a more restrictive one; and a Responsibility Advisor, which is the human-robot interaction part of the system.

Azad-Manjiri [35] develops an architecture for a deontological system constrained by Beauchamp and Childress's biomedical principles. The system determines its actions on the basis of said principles and a decision tree algorithm trained with expert ethicist judgments in a variety of cases from the biomedical domain. Building on early work by Ganascia [186], **Tufiş and Ganascia [495]** augment a belief-desire-intention rational agent model with normative constraints. They devote particular attention to the problem arising from the acquisition of new norms, which frequently stand in conflict with existing ones (for an alternative approach building on the belief-desire-intention model, see Honarvar and Ghasem-Aghaee [240, 241] discussed below).

A.3.4 Hybrid: Unspecified Hierarchy Yilmaz et al. [540] survey the field of machine ethics and propose a coherence-driven reflective equilibrium model (cf. [409]), by aid of which conflicts across heterogenous interests and values can be resolved. **Honarvar and Ghasem-Aghaee [240]** build a belief-desire-intention agent model whose decisions are based on a number of weighted features drawn from hedonic act utilitarianism (e.g., the amount of pleasure and displeasure for the agent and other parties affected by the action).

A.3.5 Ambiguous Most of the work of Saptawijaya and Pereira (c.f. [380–383, 431–433]) focuses on logic programming and prospective logic to model ethical machines. In **Han et al. [210]**, they introduce uncertainty as a factor in decision making and draw on abductive logic to accommodate it. **Madl and Franklin [320]** call for limits on ethical

machines for safety reasons. Developing on Franklin et al.'s [178] LIDA architecture—an artificial general intelligence (AGI) model of human cognition—they suggest that deliberate actions could be constrained top-down during run time, and ethical meta-rules (such as certain Kantian principles) could be implemented on a metacognitive level. Rather than start from a complete set of rules, the latter can gradually expand. The approach is exemplified by *CareBot*, an assistive simulated bot for the home care domain. **Wallach et al. [515]** also discuss the LIDA model. They demonstrate how emotions can be integrated into a LIDA-based account of the human decision making process and extend the approach to artificial moral agents.

List of Figures

1.1 Iterative and incremental design (IID) cycle, adapted from Basil and Turner [45]. See [290] for the history of the model. 5
1.2 Usage of Wizard of Oz at various stages of system design, adapted from Dow et al. [146]. 6
1.3 Summary of the research questions presented along the IID cycle. 15
2.1 Ethical theory type ratio 46
2.2 Non-technical analysis 49
2.3 Technology analysis 50
2.4 Dimension interaction 52
3.1 The housing search interface (left-hand side), and assistance from the intelligent system (right-hand side). 60
3.2 Overview of the study workflow. 65
3.3 Boxplots representing trust scores (*x-axis*) per session, across each experimental group (y-axis). 72
5.1 Screenshot of a decision during the simulation for a human-on-the-loop setting. For the decisions, the left part of the screen showed the possible crash sights, while the upper right corner showed the expert's opinion. Participants had to select their choice in the bottom right. 105
5.2 Overview of the experimental setup. Blue boxes indicate the independent variables: decision type (human-in-the-loop vs. human-on-the-loop) and expert type (human vs. AI expert). Green boxes and terms are control variables. Red italic terms are the dependent variables: the trust participants report, the responsibility they assign, and the reliance they show in the decisions they make. 108
5.3 The first two columns show the responsibility assigned in the human expert setting, the final four show the responsibility scores for the AI expert setting. A responsibility score of 1 indicates the participant thought the entity to be 'not responsible at all', while 7 implies they found them to be 'very responsible'. 111

List of Tables

1.1 overview of research questions and hypotheses covered in this thesis. 16
1.2 The contributions per chapter are classified according to Elsevier's *Contributor Roles Taxonomy*. 20
2.1 Ethical theory types taxonomy dimension 33
2.2 High-level overview to ethics categories in the context of ethical machine implementation 34
2.3 Non-technical taxonomy dimension 37
2.4 Technical taxonomy dimension. As explained in Section 2.6, "Inductive logic" is present twice. 42
2.5 Ethical theory classification. *Hybrid dominance D-C* implies both D and C are implemented, but D is dominant. The reverse is true for *Hybrid dominance C-D*. For the *Hybrid undefined dominance* the theories that are combined are noted in parentheses following the citation. 47
2.6 Non-technical dimension classification. Diversity consideration: ✓ implies yes, an empty cell implies no/not present. 48
2.7 Technical classification. ✓ implies yes/fully, ∘ implies partially, an empty cell implies no/not present. 51
3.1 Examples of `easy` and `complex` scenarios presented to users in each house search task. Each distinct constraint is colored for the benefit of the reader. 63
3.2 Hypotheses and their required comparisons. Comparisons are made either between sessions within a single group, or between different groups. Investigation of dispositional factors is not related to specific sessions or groups. 66
3.3 Number of participants per experimental condition 67
3.4 P-value results of two Fisher's exact tests on user accuracy. Difficulty (`easy`/`complex`) and system accuracy (correct/incorrect) were compared against user accuracy (correct/incorrect). 68
3.5 Ratio of participants who used the system per group by clicking the system suggestion at least once. Average usage ratio per group is shown in the last column. 68
3.6 Ratio of participants who submitted the system's suggestion after opening it. 68
3.7 Results of mixed ANOVA for average trust scores between groups. Green cells imply a significant difference between groups. The group mentioned in a green cell had a higher average trust based on Tukey's HSD test. 69

3.8 Results of mixed ANOVA for average trust scores within groups between sessions. Green cells imply a significant difference between sessions. ↗ implies trust increased between the compared sessions, ↘ indicates trust decreased. 71
3.9 Results of tested hypotheses. 72
4.1 Types of actions which cause a loss of trust: we call these failures 83
4.2 Risk Analysis of Failure leading to loss of Trust (cf. Sec. 4.3.2) 84
5.1 Scenario types 103
5.2 Both the independent variable of expert autonomy and control variables of task framing and expert order do not significantly influence the logistic model on participant reliance. 112
6.1 Summary of challenges for different aspects of HATs in aviation. 129

References

[1] Abbass, H. A. (2019). Social integration of artificial intelligence: functions, automation allocation logic and human-autonomy trust. *Cognitive Computation*, 11(2):159–171.

[2] Abel, D., MacGlashan, J., and Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In *AAAI Workshop: AI, Ethics, and Society*, volume 16, page 02, Phoenix, Arizona, USA. AAAI Press.

[3] Adhikari, S. and Basil, J. (2020). Graph database and anomaly detection based real-time, autonomous, proactive, and intelligent cyber defense for aviation. In *AIAA AVIATION 2020 FORUM*, page 2927.

[4] Adnan, N., Nordin, S. M., bin Bahruddin, M. A., and Ali, M. (2018). How trust can drive forward the user acceptance to the technology? in-vehicle technology for autonomous vehicle. *Transportation research part A: policy and practice*, 118:819–836.

[5] Akata, Z., Balliet, D., De Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., et al. (2020). A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(08):18–28.

[6] Alexander, L. and Moore, M. (2021). Deontological Ethics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition.

[7] Allen, C., Smit, I., and Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7(3):149–155.

[8] Anderson, M. and Anderson, S. L. (2008). Ethel: Toward a principled ethical eldercare robot. In *Proceedings of the AAAI Fall 2008 Symposium on AI in Eldercare: New solutions to old problems*, pages 4–11, Arlington, VA. AAAI Press.

[9] Anderson, M. and Anderson, S. L. (2010). Robot be good. *Scientific American*, 303(4):72–77.

[10] Anderson, M. and Anderson, S. L. (2018). Geneth: a general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1):337–357.

[11] Anderson, M., Anderson, S. L., and Armen, C. (2004). Towards machine ethics. In *AAAI-04 workshop on agent publishers: theory and practice, San Jose, CA*, pages 2–7, Phoenix, Arizona, USA. AAAI Press.

[12] Anderson, M., Anderson, S. L., and Armen, C. (2006). Medethex: A prototype medical ethics advisor. In *Proceedings of the 18th Conference on Innovative Applications of Artificial Intelligence - Volume 2*, IAAI'06, pages 1759–1765, Boston, Massachusetts. AAAI Press.

[13] Anderson, M., Anderson, S. L., and Berenz, V. (2019). A value-driven eldercare robot: Virtual and physical instantiations of a case-supported principle-based behavior paradigm. *Proceedings of the IEEE*, 107(3):526–540.

[14] Antun, V., Renna, F., Poon, C., Adcock, B., and Hansen, A. C. (2020). On instabilities of deep learning in image reconstruction and the potential costs of ai. *Proceedings of the National Academy of Sciences*, 117(48):30088–30095.

[15] Apiecionek, Ł., Makowski, W., Biernat, D., and Łukasik, M. (2015). Practical implementation of ai for military airplane battlefield support system. In *2015 8th International Conference on Human System Interaction (HSI)*, pages 249–253. IEEE.

[16] Araujo, T., Helberger, N., Kruikemeier, S., and De Vreese, C. H. (2020). In ai we trust? perceptions about automated decision-making by artificial intelligence. *AI & Society*, 35(3):611–623.

[17] Argall, B. D., Chernova, S., Veloso, M., and Browning, B. (2009). A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483.

[18] Arkin, R. (2007). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture. Georgia Institute of Technology GVU Technical Report GIT-GVU-07-11, S. 1–117.

[19] Arkin, R. (2008a). Governing ethical behavior: Embedding an ethical controller in a hybrid deliberative-reactive robot architecture - part ii: Formalization for ethical control. In *Proceedings of the first Artificial General Intelligence Conference 2008*, pages 51–62, Amsterdam, The Netherlands. IOS Press.

[20] Arkin, R. (2008b). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture part i: Motivation and philosophy. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 121–128, New York, NY, USA. ACM.

[21] Arkin, R. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture-part iii: Representational and architectural considerations. Technical report, Georgia Institute of Technology.

[22] Arkoudas, K., Bringsjord, S., and Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI Fall Symposium on Machine Ethics*, pages 17–23, Menlo Park, California. AAAI Press.

[23] Armstrong, S. (2015). Motivated value selection for artificial agents. In *AAAI Workshop: AI and Ethics*, volume 92, pages 12–20, Palo Alto, California. AAAI Press.

[24] Arnold, A., Dupont, G., Kobus, C., Lancelot, F., and Liu, Y.-H. (2020). Perceived usefulness of conversational agents predicts search performance in aerospace domain. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, pages 1–3. ACM.

[25] Aron, J. (2011). How innovative is apple's new voice assistant, siri?

[26] Arrow, K. J. (1972). Gifts and exchanges. *Philosophy & Public Affairs*, 1(4):343–362.

[27] Asafa, T., Afonja, T., Olaniyan, E., and Alade, H. (2018). Development of a vacuum cleaner robot. *Alexandria engineering journal*, 57(4):2911–2920.

[28] Ashley, K. D. and McLaren, B. M. (1994). A cbr knowledge representation for practical ethics. In *European Workshop on Advances in Case-Based Reasoning*, pages 180–197, Berlin, Heidelberg. Springer.

[29] Association, A. V. M. (2021). Human-animal bond. https://www.avma.org/one-health/human-animal-bond.

[30] Atkinson, K. and Bench-Capon, T. (2007). Action-based alternating transition systems for arguments about action. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, volume 7, pages 24–29, Palo Alto, California. AAAI Press.

[31] Atkinson, K. and Bench-Capon, T. (2008). Addressing moral problems through practical reasoning. *Journal of Applied Logic*, 6(2):135–151.

[32] Attenberg, J., Ipeirotis, P., and Provost, F. (2015). Beat the machine: Challenging humans to find a predictive model's "unknown unknowns". *Journal of Data and Information Quality (JDIQ)*, 6(1):1–17.

[33] Attig, C., Wessel, D., and Franke, T. (2017). Assessing personality differences in human-technology interaction: an overview of key self-report scales to predict successful interaction. In *International Conference on Human-Computer Interaction*, pages 19–29. Springer.

[34] Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., and Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729):59.

[35] Azad-Manjiri, M. (2014). A new architecture for making moral agents based on c4. 5 decision tree algorithm. *International Journal of Information Technology and Computer Science (IJITCS)*, 6(5):50.

[36] Azevedo-Sa, H., Jayaraman, S. K., Esterwood, C. T., Yang, X. J., Robert Jr, L. P., and Tilbury, D. M. (2020). Comparing the effects of false alarms and misses on humans' trust in (semi) autonomous vehicles. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 113–115.

[37] Bajones, M., Weiss, A., and Vincze, M. (2016). Help, anyone? a user study for modeling robotic behavior to mitigate malfunctions with the help of the user. *arXiv preprint arXiv:1606.02547*.

[38] Baker, A. L., Phillips, E. K., Ullman, D., and Keebler, J. R. (2018). Toward an Understanding of Trust Repair in Human-Robot Interaction: Current Research and Future Directions. *ACM Transactions on Interactive Intelligent Systems*, 8(4):1–30.

[39] Baker, A. L., Schaefer, K. E., and Hill, S. G. (2019). Teamwork and communication methods and metrics for human-autonomy teaming. Technical report, CCDC Army Research Laboratory Aberdeen Proving Ground United States.

[40] Balkmar, D. and Mellström, U. (2019). Towards an anthropology of transport affect: The place of emotions, gender, and power in smart mobilities. In *Gendering smart mobilities*, pages 57–74. Routledge.

[41] Balzer, R., Erman, L. D., London, P., and Williams, C. (1980). Hearsay-ii: A domain-independent framework for expert systems. In *AAAI*, volume 1, pages 108–110.

[42] Banks, M. R., Willoughby, L. M., and Banks, W. A. (2008). Animal-assisted therapy and loneliness in nursing homes: use of robotic versus living dogs. *Journal of the American Medical Directors Association*, 9(3):173–177.

[43] Baral, C. (2003). *Knowledge representation, reasoning and declarative problem solving*. Cambridge university press, Cambridge,UK.

[44] Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.

[45] Basil, V. R. and Turner, A. J. (1975). Iterative enhancement: A practical technique for software development. *IEEE Transactions on Software Engineering*, SE-1(4):390–396.

[46] Battiste, V., Lachter, J., Brandt, S., Alvarez, A., Strybel, T. Z., and Vu, K.-P. L. (2018). Human-Automation Teaming: Lessons Learned and Future Directions. In Yamamoto, S. and Mori, H., editors, *Human Interface and the Management of Information. Information in Applications and Services*, Lecture Notes in Computer Science, pages 479–493, Cham. Springer International Publishing.

[47] Baum, S. D. (2017). Social choice ethics in artificial intelligence. *AI & Society*, 32:1–12.

[48] Bauman, C. W., McGraw, A. P., Bartels, D. M., and Warren, C. (2014). Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass*, 8(9):536–554.

[49] Beauchamp, T. L. and Childress, J. F. (2001). *Principles of biomedical ethics*. Oxford University Press, USA, New York, NY, USA.

[50] Beggiato, M. and Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation research part F: traffic psychology and behaviour*, 18:47–57.

[51] Benbasat, I. and Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the association for information systems*, 6(3):4.

[52] Bench-Capon, T. (2020). Ethical approaches and autonomous systems. *Artificial Intelligence*, 281:1–15.

[53] Bench-Capon, T., Atkinson, K., and Chorley, A. (2005). Persuasion and value in legal argument. *Journal of Logic and Computation*, 15(6):1075–1097.

[54] Bench-Capon, T. and Sartor, G. (2003). A model of legal reasoning with cases incorporating theories and values. *Artificial Intelligence*, 150(1-2):97–143.

[55] Bendel, O. (2019). *Handbuch Maschinenethik*. Springer.

[56] Bente, G., Eschenburg, F., and Aelker, L. (2007). Effects of simulated gaze on social presence, person perception and personality attribution in avatar-mediated communication. In *Presence 2007: Proceedings of the 10th Annual International Workshop on Presence, October 25-27, 2007, Barcelona, Spain*, pages 207–14.

[57] Bentham, J. (1996). *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation*. Clarendon Press.

[58] Berger, B., Adam, M., Rühr, A., and Benlian, A. (2021). Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, 63(1):55–68.

[59] Berreby, F., Bourgne, G., and Ganascia, J.-G. (2015). Modelling moral reasoning and ethical responsibility with logic programming. In *Logic for Programming, Artificial Intelligence, and Reasoning*, pages 532–548, Berlin, Heidelberg. Springer.

[60] Berreby, F., Bourgne, G., and Ganascia, J.-G. (2017). A declarative modular framework for representing and applying ethical principles. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 96–104, Richland, SC. International Foundation for Autonomous Agents and Multiagent Systems.

[61] Billings, D. R., Schaefer, K. E., Chen, J. Y., and Hancock, P. A. (2012). Human-robot interaction: developing trust in robots. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction - HRI '12*, page 109, Boston, Massachusetts, USA. ACM Press.

[62] Blackburn, S. (2002). *Being good: A short introduction to ethics*. OUP Oxford, Oxford, United Kingdom.

[63] Blackburn, S. (2016). *The Oxford Dictionary of Philosophy*. Oxford University Press, Oxford, UK.

[64] Boehm, B. W. (1991). Software risk management: principles and practices. *IEEE software*, 8(1):32–41.

[65] Bogg, A., Birrell, S., Bromfield, M. A., and Parkes, A. M. (2020). Can we talk? How a talking agent can improve human autonomy team performance. *Theoretical Issues in Ergonomics Science*, 22(4):1–22.

[66] Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293):1573–1576.

[67] Bonnemains, V., Saurel, C., and Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology*, 20(1):41–58.

[68] Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, Oxford,UK.

[69] Bourget, D. and Chalmers, D. J. (2014). What do philosophers believe? *Philosophical studies*, 170(3):465–500.

[70] Brandt, S. L., Lachter, J., Russell, R., and Shively, R. J. (2017). A human-autonomy teaming approach for a flight-following task. In *International Conference on Applied Human Factors and Ergonomics*, pages 12–22. Springer.

[71] Braun, M., Mainz, A., Chadowitz, R., Pfleging, B., and Alt, F. (2019). At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–11, New York, NY, USA. Association for Computing Machinery.

[72] Braun, M., Völkel, S. T., Hussmann, H., Frison, A.-K., Alt, F., and Riener, A. (2018). Beyond transportation: How to keep users attached when they are neither driving nor owning automated cars? In *Adjunct Proceedings of the 10th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 175–180.

[73] Brewer, R. N., Findlater, L., Kaye, J., Lasecki, W., Munteanu, C., and Weber, A. (2018). Accessible voice interfaces. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 441–446.

[74] Bringsjord, S. and Taylor, J. (2012). *Robot ethics: the ethical and social implication of robotics*, chapter Introducing divine-command robot ethics, pages 85–108. MIT Press, Cambridge, Massachusetts, USA.

[75] Brooks, D. J., Begum, M., and Yanco, H. A. (2016). Analysis of reactions towards failures and recovery strategies for autonomous robots. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 487–492, New York, NY,USA. IEEE.

[76] Brscić, D., Kidokoro, H., Suehiro, Y., and Kanda, T. (2015). Escaping from children's abuse of social robots. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '15, pages 59–66, New York, NY, USA. ACM.

[77] Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental and Theoretical Artificial Intelligence*, 26(3):355–372.

[78] Bryson, J. J. (2010). Robots should be slaves. *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues*, 8:63–74.

[79] Burke, R., Felfernig, A., and Göker, M. H. (2011). Recommender systems: An overview. *Ai Magazine*, 32(3):13–18.

[80] Burton, J. W., Stein, M.-K., and Jensen, T. B. (2020). A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239.

[81] Cakmak, M. and Takayama, L. (2014). Teaching people how to teach robots: The effect of instructional materials and dialog design. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 431–438, Toronto, Canada. ACM.

[82] Calhoun, G., Ruff, H., Behymer, K., and Frost, E. (2018). Human-autonomy teaming interface design considerations for multi-unmanned vehicle control. *Theoretical issues in ergonomics science*, 19(3):321–352.

[83] Cambria, E., Das, D., Bandyopadhyay, S., and Feraco, A. (2017). Affective computing and sentiment analysis. In *A practical guide to sentiment analysis*, pages 1–10. Springer.

[84] Carim, G. C., Saurin, T. A., Havinga, J., Rae, A., Dekker, S. W., and Éder Henriqson (2016). Using a procedure doesn't mean following it: A cognitive systems approach to how a cockpit manages emergencies. *Safety Science*, 89:147–157.

[85] Carlson, J. and Murphy, R. R. (2005). How ugvs physically fail in the field. *IEEE Transactions on robotics*, 21(3):423–437.

[86] Castelfranchi, C. and Falcone, R. (1998). Principles of trust for MAS: cognitive anatomy, social importance, and quantification. In Demazeau, Y., editor, *Proceedings of the Third International Conference on Multiagent Systems, ICMAS 1998, Paris, France, July 3-7, 1998*, pages 72–79, New York, NY,USA. IEEE.

[87] Castelo, N., Bos, M. W., and Lehmann, D. R. (2019). Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825.

[88] Cavazos, J. G., Phillips, P. J., Castillo, C. D., and O'Toole, A. J. (2020). Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science*, 3(1):101–111.

[89] Cave, S. and Dihal, K. (2019). Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence*, 1(2):74–78.

[90] Cervantes, J.-A., Rodríguez, L.-F., López, S., and Ramos, F. (2013). A biologically inspired computational model of moral decision making for autonomous agents. In *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2013 12th IEEE International Conference on*, pages 111–117, New York, NY, USA. IEEE.

[91] Cervantes, J.-A., Rodríguez, L.-F., López, S., Ramos, F., and Robles, F. (2016). Autonomous agents and ethical decision-making. *Cognitive Computation*, 8(2):278–296.

[92] Cesafsky, L., Stayton, E., and Cefkin, M. (2019). Calibrating Agency: Human-Autonomy Teaming and the Future of Work amid Highly Automated Systems. *Ethnographic Praxis in Industry Conference Proceedings*, 2019(1):65–82.

[93] Chakraborti, T., Kambhampati, S., Scheutz, M., and Zhang, Y. (2017). Ai challenges in human-robot cognitive teaming. *arXiv preprint arXiv:1707.04775*.

[94] Chan, H. S., Shan, H., Dahoun, T., Vogel, H., and Yuan, S. (2019). Advancing drug discovery via artificial intelligence. *Trends in pharmacological sciences*, 40(8):592–604.

[95] Chancey, E. T., Bliss, J. P., Yamani, Y., and Handley, H. A. (2017). Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human factors*, 59(3):333–345.

[96] Chen, J., Chen, C., B. Walther, J., and Sundar, S. S. (2021). Do you feel special when an ai doctor remembers you? individuation effects of ai vs. human doctors on user experience. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA. Association for Computing Machinery.

[97] Chen, J. Y. and Barnes, M. J. (2014). Human–agent teaming for multirobot control: A review of human factors issues. *IEEE Transactions on Human-Machine Systems*, 44(1):13–29.

[98] Chen, J. Y., Barnes, M. J., Selkowitz, A. R., and Stowers, K. (2016a). Effects of agent transparency on human-autonomy teaming effectiveness. In *2016 IEEE international conference on Systems, man, and cybernetics (SMC)*, pages 001838–001843. IEEE.

[99] Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., and Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, 19(3):259–282.

[100] Chen, J. Y. C. (2018). Human-autonomy teaming in military settings. *Theoretical Issues in Ergonomics Science*, 19(3):255–258.

[101] Chen, Y., Argentinis, J. E., and Weber, G. (2016b). Ibm watson: how cognitive computing can be applied to big data challenges in life sciences research. *Clinical therapeutics*, 38(4):688–701.

[102] Chiang, C.-W. and Yin, M. (2021). You'd better stop! understanding human reliance on machine learning models under covariate shift. In *13th ACM Web Science Conference 2021*, pages 120–129.

[103] Christen, M., Narvaez, D., Zenk, J. D., Villano, M., Crowell, C. R., and Moore, D. R. (2021). Trolley dilemma in the sky: Context matters when civilians and cadets make remotely piloted aircraft decisions. *PLoS one*, 16(3):e0247273.

[104] Cialdini, R. B. (1993). *Influence: The psychology of persuasion*. Morrow New York, New York, NY, USA.

[105] Clarke, E., Grumberg, O., and Peled, D. A. (2000). *Model Checking*. MIT Press, Cambridge,MA,USA.

[106] Cloos, C. (2005). The Utilibot Project: An Autonomous Mobile Robot Based on Utilitarianism. In *Machine Ethics: Papers from the 2005 AAAI Fall Symposium*, pages 38–45, Menlo Park, CA, USA. AAAI Press.

[107] Cockburn, A., Gutwin, C., Palanque, P., Deleris, Y., Trask, C., Coveney, A., Yung, M., and MacLean, K. (2017). Turbulent touch: Touchscreen input for cockpit flight displays. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 6742–6753. ACM.

[108] Coghlan, S. (2020). *Ethical dimensions of animal-assisted interventions*, pages 69–97. Nova Science Publishers.

[109] Coghlan, S., Vetere, F., Waycott, J., and Neves, B. B. (2019). Could social robots make us kinder or crueller to humans and animals? *International Journal of Social Robotics*, 11(5):741–751.

[110] Cointe, N., Bonnet, G., and Boissier, O. (2016). Ethical judgment of agents' behaviors in multi-agent systems. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 1106–1114, Richland, SC, USA. International Foundation for Autonomous Agents and Multiagent Systems.

[111] Coleman, K. G. (2001). Android arete: Toward a virtue ethic for computational agents. *Ethics and Information Technology*, 3(4):247–265.

[112] Constant, B. (2013). *Des réactions politiques*. Presses Électroniques de France, Paris,France.

[113] Copp, D. (2005). *The Oxford handbook of ethical theory*. Oxford University Press, Oxford, United Kingdom.

[114] Cover, M., Reichlen, C., Matessa, M., and Schnell, T. (2018a). Analysis of Airline Pilots Subjective Feedback to Human Autonomy Teaming in a Reduced Crew Environment. In Yamamoto, S. and Mori, H., editors, *Human Interface and the Management of Information. Information in Applications and Services*, Lecture Notes in Computer Science, pages 359–368, Cham. Springer International Publishing.

[115] Cover, M., Reichlen, C., Matessa, M., and Schnell, T. (2018b). Analysis of airline pilots subjective feedback to human autonomy teaming in a reduced crew environment. In *International Conference on Human Interface and the Management of Information*, pages 359–368. Springer.

[116] Cox, K. B. (2003). The Effects of Intrapersonal, Intragroup, and Intergroup Conflict on Team Performance Effectiveness and Work Satisfaction. *Nursing Administration Quarterly*, 27(2):153–163.

[117] Crespi, V., Galstyan, A., and Lerman, K. (2008). Top-down vs bottom-up methodologies in multi-agent system design. *Autonomous Robots*, 24(3):303–313.

[118] Crook, J. N., Edelman, D. B., and Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3):1447–1465.

[119] Crosby, M., Beyret, B., Shanahan, M., Hernández-Orallo, J., Cheke, L., and Halina, M. (2020). The animal-ai testbed and competition. In *NeurIPS 2019 Competition and Demonstration Track*, pages 164–176. PMLR.

[120] Cummings, M. (2017). Artificial intelligence and the future of warfare. Technical report, International Security Department and US and the Americas Program.

[121] Cummings, M., Huang, L., Zhu, H., Finkelstein, D., and Wei, R. (2019). The impact of increasing autonomy on training requirements in a UAV supervisory control task. *Journal of Cognitive Engineering and Decision Making*, 13(4):295–309.

[122] Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies—why and how. *Knowledge-based systems*, 6(4):258–266.

[123] Danaher, J. (2020). Welcoming robots into the moral circle: a defence of ethical behaviourism. *Science and Engineering Ethics*, 26(4):2023–2049.

[124] Dancy, J. (1999). Can a particularist learn the difference between right and wrong? In *The proceedings of the twentieth world congress of philosophy*, volume 1, pages 59–72, Charlottesville, Virginia,US. Philosophy Documentation Center.

[125] Dancy, J. (2000). The particularist's progress. In *Moral Particularism*. Oxford University Press, Oxford,UK.

[126] Dancy, J. et al. (2004). *Ethics without principles*. Oxford University Press on Demand, Oxford,UK.

[127] Darling, K., Nandy, P., and Breazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. In *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pages 770–775. IEEE.

[128] Davenport, D. (2014). Moral mechanisms. *Philosophy & Technology*, 27(1):47–60.

[129] de Lima, Y. O. and de Souza, J. M. (2017). The future of work: Insights for cscw. In *2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 42–47. IEEE.

[130] de Visser, E. J., Pak, R., and Shaw, T. H. (2018). From 'automation'to 'autonomy': the importance of trust repair in human–machine interaction. *Ergonomics*, 61(10):1409–1427.

[131] De Waal, F. B. (1999). Anthropomorphism and anthropodenial: consistency in our thinking about humans and other animals. *Philosophical Topics*, 27(1):255–280.

[132] de Waal, F. B. (2011). What is an animal emotion?: What is an animal emotion? *Annals of the New York Academy of Sciences*, 1224(1):191–206.

[133] DeChurch, L., Mesmer-Magnus, J., and Doty, D. (2013). Moving Beyond Relationship and Task Conflict: Toward a Process-State Perspective. *Journal of Applied Psychology*, 98(4):559–578.

[134] Dehghani, M., Tomai, E., Forbus, K. D., and Klenk, M. (2008). An integrated reasoning approach to moral decision-making. In *AAAI*, pages 1280–1286, Chicago, Illinois, USA. AAAI Press.

[135] Del Duchetto, F., Kucukyilmaz, A., Iocchi, L., Hanheide, M., Duchetto, F. D., Kucukyilmaz, A., Iocchi, L., and Hanheide, M. (2018). Do Not Make the Same Mistakes Again and Again: Learning Local Recovery Policies for Navigation From Human Demonstrations. *IEEE Robotics and Automation Letters*, 3(4):4084–4091.

[136] Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., and Ebel, P. (2021). The future of human-ai collaboration: a taxonomy of design knowledge for hybrid intelligence systems. *arXiv preprint arXiv:2105.03354*.

[137] Demir, M., McNeese, N. J., and Cooke, N. J. (2017). Team situation awareness within the context of human-autonomy teaming. *Cognitive Systems Research*, 46:3–12.

[138] Dennett, D. C. (1989). *The intentional stance*. MIT press, Cambridge,MA,USA.

[139] Dennis, L., Fisher, M., Slavkovik, M., and Webster, M. (2016). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, 77:1–14.

[140] Dennis, L. A., Fisher, M., and Winfield, A. F. (2015). Towards verifiably ethical robot behaviour. In *AAAI Workshop: AI and Ethics*, pages 45–52, Palo Alto, California, US. AAAI Press.

[141] Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., and Yanco, H. (2013). Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*, pages 251–258, Tokyo, Japan. IEEE Press.

[142] Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114.

[143] Doherty, P. and Rudol, P. (2007). A uav search and rescue scenario with human body detection and geolocalization. In *Australasian Joint Conference on Artificial Intelligence*, pages 1–13. Springer.

[144] Dominguez-Péry, C. and Vuddaraju, L. N. R. (2020). From Human Automation Interactions to Social Human Autonomy Machine Teaming in Maritime Transportation. In Sharma, S. K., Dwivedi, Y. K., Metri, B., and Rana, N. P., editors, *Re-imagining Diffusion and Adoption of Information Technology and Systems: A Continuing Conversation*, IFIP Advances in Information and Communication Technology, pages 45–56, Cham. Springer International Publishing.

[145] Dong, Y., Ai, J., and Liu, J. (2019). Guidance and control for own aircraft in the autonomous air combat: A historical review and future prospects. *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, 233(16):5943–5991.

[146] Dow, S., MacIntyre, B., Lee, J., Oezbek, C., Bolter, J. D., and Gandy, M. (2005). Wizard of oz support throughout an iterative design process. *IEEE Pervasive Computing*, 4(4):18–26.

[147] Driscoll, C. J. (2020). *Animal-assisted interventions for health and human service professionals*. Nova Science Publishers.

[148] Dubey, A., Abhinav, K., Jain, S., Arora, V., and Puttaveerana, A. (2020). Haco: A framework for developing human-ai teaming. In *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly known as India Software Engineering Conference*, pages 1–9.

[149] Dudhwala, F. and Larsen, L. B. (2019). Recalibration in counting and accounting practices: Dealing with algorithmic output in public and private. *Big Data & Society*, 6(2):2053951719858751.

[150] ECA (2017). Airline pilot profession code of ethics. https://www.talpa.org/wp-content/uploads/2019/07/Future_Airline_Pilot_Profession_Pillar-I_Code_of_Ethics_FINAL.pdf. European Cockpit Association.

[151] Eduard Brandstaetter, G. G. and Hertwig, R. (2006). The priority heuristic: Making choices without trade-offs. *Psychological Review*, 113(2):409–462.

[152] Eisenhardt, K. M. (1989). Building theories from case study research. *The Academy of Management Review*, 14(4):532–550.

[153] Elish, M. C. (2019). Moral crumple zones: Cautionary tales in human-robot interaction (pre-print). *Engaging Science, Technology, and Society (pre-print)*.

[154] Endsley, M. R. (1995). Measurement of Situation Awareness in Dynamic Systems. *Human Factors*, 37(1):65–84. Publisher: SAGE Publications Inc.

[155] Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human factors*, 59(1):5–27.

[156] Epstein, Z., Levine, S., Rand, D. G., and Rahwan, I. (2020). Who gets credit for ai-generated art? *Iscience*, 23(9):101515.

[157] Erlei, A., Nekdem, F., Meub, L., Anand, A., and Gadiraju, U. (2020). Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 43–52.

[158] Etzioni, A. and Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4):403–418.

[159] Evins, R., Vaidyanathan, R., and Burgess, S. (2014). Multi-material compositional pattern-producing networks for form optimisation. In *European Conference on the Applications of Evolutionary Computation*, pages 189–200, Berlin, Heidelberg. Springer.

[160] Falcone, R. and Castelfranchi, C. (2001). Social trust: A cognitive approach. In Castelfranchi, C. and Tan, Y.-H., editors, *Trust and Deception in Virtual Societies*, pages 55–90. Springer, Dordrecht.

[161] Fank, J., Richardson, N. T., and Diermeyer, F. (2019). Anthropomorphising driver-truck interaction: a study on the current state of research and the introduction of two innovative concepts. *Journal on Multimodal User Interfaces*, 13(2):99–117.

[162] Fast, E. and Horvitz, E. (2017). Long-term trends in the public perception of artificial intelligence. In *Thirty-First AAAI Conference on Artificial Intelligence*, pages 963–969, Phoenix, Arizona, USA. AAAI Press.

[163] Feinberg, J. (1987). *The moral limits of the criminal law. 1, Harm to others*. Oxford University Press, New York, NY,USA.

[164] Ferguson, J. M. (2014). Terminally haunted: Aviation ghosts, hybrid buddhist practices, and disaster aversion strategies amongst airport workers in myanmar and thailand. *The Asia Pacific Journal of Anthropology*, 15(1):47–64.

[165] Ferguson, J. M. and Ayuttacorn, A. (2019). Air male: Exploring flight attendant masculinities in north america and thailand. *The Asia Pacific Journal of Anthropology*, 20(4):328–343.

[166] Feroz, I., Ahmad, N., Iqbal, M. W., Main, N. A., and Shahzad, S. K. (2019). People perception of autonomous vehicles: Legal and ethical issues. *International Journal of Advanced and Applied Sciences*, 6(5):92–101.

[167] Fessler, D. M., Barrett, H. C., Kanovsky, M., Stich, S., Holbrook, C., Henrich, J., Bolyanatz, A. H., Gervais, M. M., Gurven, M., Kushnick, G., et al. (2015). Moral parochialism and contextual contingency across seven societies. *Proceedings of the Royal Society B: Biological Sciences*, 282(1813):20150907.

[168] Feyerherm, A. E. and Rice, C. L. (2002). Emotional intelligence and team performance: The good, the bad and the ugly. *The International Journal of Organizational Analysis*.

[169] Fikes, R. E. and Nilsson, N. J. (1971). Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, 2:189–208.

[170] Fine, A. H. (2010). *Handbook on animal-assisted therapy: Theoretical foundations and guidelines for practice*. academic press.

[171] Fischer, J. E., Greenhalgh, C., Jiang, W., Ramchurn, S. D., Wu, F., and Rodden, T. (2021). In-the-loop or on-the-loop? interactional arrangements to support team coordination with a planning agent. *Concurrency and Computation: Practice and Experience*, 33(8):e4082. e4082 cpe.4082.

[172] Fisher, M. (2011). *An Introduction to Practical Formal Methods Using Temporal Logic*. Wiley, New York, USA.

[173] Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, 14(3):349–379.

[174] Forsyth, D. R., Zyzniewski, L. E., and Giammanco, C. A. (2002). Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin*, 28(1):54–65.

[175] Fox, M., Long, D., and Magazzeni, D. (2017). Explainable planning. In *IJCAI-17 Workshop on Explainable AI*, Melbourne, Australia. IJCAI.

[176] Franke, T., Attig, C., and Wessel, D. (2019). A personal resource for technology interaction: development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467.

[177] Frankfurt, H. (1994). An alleged asymmetry between actions and omissions. *Ethics*, 104(3):620–623.

[178] Franklin, S. and Patterson Jr, F. (2006). The lida architecture: Adding new modes of learning to an intelligent, autonomous, software agent. *pat*, 703:764–1004.

[179] Frazier, M. L., Johnson, P. D., and Fainshmidt, S. (2013). Development and validation of a propensity to trust scale. *Journal of Trust Research*, 3(2):76–97.

[180] Freier, N. G., Nelson, E. J., Rotondo, A., and Kong, W. K. (2009). *The Moral Accountability of a Personified Agent: Young Adults' Conceptions*, page 4609–4614. Association for Computing Machinery, New York, NY, USA.

[181] Freude, H., Heger, O., and Niehaves, B. (2019). Unveiling emotions: Attitudes toward affective technology. In *ICIS 2019 Proceedings*.

[182] Furbach, U., Schon, C., and Stolzenburg, F. (2014). Automated reasoning in deontic logic. In *International Workshop on Multi-disciplinary Trends in Artificial Intelligence*, pages 57–68, Cham, Switzerland. Springer.

[183] Gabriel, I. (2020). Artificial intelligence, values and alignment. *arXiv preprint arXiv:2001.09768*.

[184] Gainer, P., Dixon, C., Dautenhahn, K., Fisher, M., Hustadt, U., Saunders, J., and Webster, M. (2017). Cruton: Automatic verification of a robotic assistant's behaviours. In Petrucci, L., Seceleanu, C., and Cavalcanti, A., editors, *Critical Systems: Formal Methods and Automated Verification - Joint 22nd International Workshop on Formal Methods for Industrial Critical Systems - and - 17th International Workshop on Automated Verification of Critical Systems, FMICS-AVoCS 2017, Turin, Italy, September 18-20, 2017, Proceedings*, volume 10471 of *Lecture Notes in Computer Science*, pages 119–133, Turin, Italy. Springer.

[185] Ganascia, J.-G. (2007). Ethical system formalization using non-monotonic logics. In *Proceedings of the Cognitive Science conference*, volume 29, pages 1013–1018, Nashville, US. Cognitive Science Society.

[186] Ganascia, J.-G. (2012). An agent-based formalization for resolving ethical conflicts. In *Belief change, Non-monotonic reasoning and Conflict resolution Workshop-ECAI, Montpellier, France,(August 2012)*, pages 1–7, Amsterdam,NL. IOD Press.

[187] Gauci, J., Xuereb, M., Muscat, A., and Zammit-Mangion, D. (2017). Multi-modal interaction between pilots and avionic systems on-board large commercial aircraft. In

Harris, D., editor, *Engineering Psychology and Cognitive Ergonomics: Cognition and Design*, volume 10276, pages 200–210. Springer International Publishing.

[188] Gerling, K., Hebesberger, D., Dondrup, C., Körtner, T., and Hanheide, M. (2016). Robot deployment in long-term care. *Zeitschrift für Gerontologie und Geriatrie*, 49(4):288–297.

[189] Göbelbecker, M., Keller, T., Eyerich, P., Brenner, M., and Nebel, B. (2010). Coming up with good excuses: What to do when no plan can be found. In *Proceedings of the 20th International Conference on Automated Planning and Scheduling (ICAPS 2010)*, pages 81–88, Toronto, Canada. AAAI Press.

[190] Gockley, R., Bruce, A., Forlizzi, J., Michalowski, M., Mundell, A., Rosenthal, S., Sellner, B., Simmons, R., Snipes, K., Schultz, A. C., et al. (2005). Designing robots for long-term social interaction. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1338–1343, New York, NY,USA. IEEE.

[191] Goodman, B. and Flaxman, S. (2017). European union regulations on algorithmic decision making and a" right to explanation". *AI Magazine*, 38(3):50–57.

[192] Gorman, J. C., Cooke, N. J., Pedersen, H. K., Winner, J., Andrews, D., and Amazeen, P. G. (2006). Changes in team composition after a break: building adapative command-and-control teams. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 50, pages 487–491. SAGE Publications Sage CA: Los Angeles, CA.

[193] Gosper, S., Trippas, J., Richards, H., Allison, F., Sear, C., Khorasani, S., and Mattioli, F. (2021). Understanding the utility of digital flight assistants: A preliminary analysis. In *Proceedings of the 3st International Conference on Conversational User Interfaces*, New York, NY, USA. Association for Computing Machinery.

[194] Govindarajulu, N. S. and Bringsjord, S. (2017). On automating the doctrine of double effect. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4722–4730, Melbourne, Australia. International Joint Conferences on Artificial Intelligence.

[195] Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018). When will ai exceed human performance? evidence from ai experts. *Journal of Artificial Intelligence Research*, 62:729–754.

[196] Gregg, A. (2020). In a first, air force uses ai on military jet. The Washington Post, published online.

[197] Grgić-Hlača, N., Engel, C., and Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

[198] Groom, V., Srinivasan, V., Bethel, C. L., Murphy, R., Dole, L., and Nass, C. (2011). Responses to robot social roles and social role framing. In *2011 International Conference on Collaboration Technologies and Systems (CTS)*, pages 194–203, New York, NY,USA. IEEE.

[199] Guarini, M. (2006). Particularism and the classification and reclassification of moral cases. *IEEE Intelligent Systems*, 21(4):22–28.

[200] Guarini, M. (2012). Moral cases, moral reasons, and simulation. *AISB/IACAP World Congr*, 21(4):22–28.

[201] Gudex, C., Kind, P., et al. (1988). The qaly toolkit. Technical report, Centre for Health Economics, University of York.

[202] Gupta, R., Kurtz, Z. T., Scherer, S., and Smereka, J. M. (2018). Open Problems in Robotic Anomaly Detection. *CoRR*, abs/1809.0.

[203] Haber, J. G. (1993). *Doing and Being, Selected Readings in Moral Philosophy*, volume 208. Macmillan Publishing Co., London, UK.

[204] Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., Enk, A., et al. (2018). Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology*, 29(8):1836–1842.

[205] Hafner, C. D. and Berman, D. H. (2002). The role of context in case-based legal reasoning: teleological, temporal, and procedural. *Artificial Intelligence and Law*, 10(1-3):19–64.

[206] Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30(1):99–120.

[207] Haimson, C., Paul, C. L., Joseph, S., Rohrer, R., and Nebesh, B. (2019). Do We Need "Teaming" to Team with a Machine? In Schmorrow, D. D. and Fidopiastis, C. M., editors, *Augmented Cognition*, Lecture Notes in Computer Science, pages 169–178, Cham. Springer International Publishing.

[208] Halpern, J. Y. and Vardi, M. Y. (1991). Model checking vs. theorem proving: a manifesto. *Artificial intelligence and mathematical theory of computation*, 212:151–176.

[209] Hamet, P. and Tremblay, J. (2017). Artificial intelligence in medicine. *Metabolism*, 69:S36–S40.

[210] Han, T., Saptawijaya, A., and Moniz Pereira, L. (2012). Moral reasoning under uncertainty. In *Logic for Programming, Artificial Intelligence, and Reasoning*, pages 212–227, Berlin, Heidelberg. Springer.

[211] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(5):517–527.

[212] Hanheide, M., Göbelbecker, M., Horn, G. S., Pronobis, A., Sjöö, K., Aydemir, A., Jensfelt, P., Gretton, C., Dearden, R., Janicek, M., Zender, H., Kruijff, G.-J., Hawes, N., and Wyatt, J. L. (2017a). Robot task planning and explanation in open and uncertain worlds. *Artificial Intelligence*, 247:119–150.

[213] Hanheide, M., Hebesberger, D., Krajník, T., Krajnik, T., and Others (2017b). The When, Where, and How: An Adaptive Robotic Info-Terminal for Care Home Residents.

In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pages 341–349, New York, New York, USA. ACM/IEEE, ACM Press.

[214] Hardin, R. (2006). *Trust*. Cambridge: Polity.

[215] Hare, R. M. and Hare, R. (1963). *Freedom and reason*, volume 92. Oxford Paperbacks, Oxford, UK.

[216] Harman, G. (2005). Moral particularism and transduction. *Philosophical Issues*, 15:44–55.

[217] Harman, G. and Thomson, J. J. (1996). *Moral Relativism*, pages 3–64. Blackwell Publishers, Cambridge, MA, USA.

[218] Hawes, N., Burbridge, C., Jovan, F., Kunze, L., Lacerda, B., Mudrová, L., Young, J., Wyatt, J., Hebesberger, D., Körtner, T., others, Ambrus, R., Bore, N., Folkesson, J., Jensfelt, P., Beyer, L., Hermans, A., Leibe, B., Aldoma, A., Fäulhammer, T., Zillich, M., Vincze, M., Al-Omari, M., Chinellato, E., Duckworth, P., Gatsoulis, Y., Hogg, D. C., Cohn, A. G., Dondrup, C., Fentanes, J. P., Krajník, T., Santos, J. M., Duckett, T., and Hanheide, M. (2017). The STRANDS Project: Long-Term Autonomy in Everyday Environments. *Robotics and Automation Magazine*.

[219] Hebesberger, D., Koertner, T., Gisinger, C., Pripfl, J., and Dondrup, C. (2016). Lessons learned from the deployment of a long-term autonomous robot as companion in physical therapy for older adults with dementia a mixed methods study. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 27–34, Christchurch, New Zealand. IEEE.

[220] Hebesberger, D. V., Dondrup, C., Gisinger, C., and Hanheide, M. (2017). Patterns of Use: How Older Adults with Progressed Dementia Interact with a Robot. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, pages 131–132, New York, New York, USA. ACM, ACM Press.

[221] Hedlund, M. and Persson, E. (2021). Expert responsibility in ai development. *Proceedings of the International Conference of Public Policy (ICPP5)*.

[222] Heerink, M., Krose, B., Evers, V., and Wielinga, B. (2006). The influence of a robot's social abilities on acceptance by elderly users. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pages 521–526, New York, NY,USA. IEEE.

[223] Helmreich, R. L. and Davies, J. M. (2004). Culture, threat, and error: lessons from aviation. *Canadian Journal of Anesthesia*, 51(1):R1–R4.

[224] Hendee, J. C. (1970). An expert system for marine environmental monitoring in the florida keys national marine sanctuary and florida bay. *WIT Transactions on Ecology and the Environment*, 25.

[225] Herkert, J., Borenstein, J., and Miller, K. (2020). The boeing 737 max: Lessons for engineering ethics. *Science and engineering ethics*, 26(6):2957–2974.

[226] Herrero, R. P., Fentanes, J. P., and Hanheide, M. (2018). Getting to Know Your Robot Customers: Automated Analysis of User Identity and Demographics for Robots in the Wild. *IEEE Robotics and Automation Letters*, 3(4):3733–3740.

[227] Herzig, A., Lorini, E., Hübner, J. F., and Vercouter, L. (2010). A logic of trust and reputation. *Logic Journal of the IGPL*, 18(1):214–244.

[228] Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and information technology*, 16(3):197–206.

[229] Hidalgo, C. A., Orghian, D., Canals, J. A., De Almeida, F., and Martin, N. (2021). *How humans judge machines*. MIT Press.

[230] Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology*, 11(1):19–29.

[231] Ho, G., Wheatley, D., and Scialfa, C. T. (2005). Age differences in trust and reliance of a medication management system. *Interacting with Computers*, 17(6):690–710.

[232] Hobbes, T. (1980). Leviathan (1651). *Glasgow 1974*.

[233] Hoff, K. A. and Bashir, M. (2015a). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, 57(3):407–434.

[234] Hoff, K. A. and Bashir, M. (2015b). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 57(3):407–434. PMID: 25875432.

[235] Hoffman, R. R., Johnson, M., Bradshaw, J. M., and Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, 28(1):84–88.

[236] Holford, W. D. (2020). An ethical inquiry of the effect of cockpit automation on the responsibilities of airline pilots: Dissonance or meaningful control? *Journal of Business Ethics*, pages 1–17.

[237] Holliday, D., Wilson, S., and Stumpf, S. (2016). User trust in intelligent systems: A journey over time. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 164–168.

[238] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

[239] Holt, C. A. and Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, 92(5):1644–1655.

[240] Honarvar, A. R. and Ghasem-Aghaee, N. (2009a). An artificial neural network approach for creating an ethical artificial agent. In *Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on*, pages 290–295, Daejeon, South Korea. IEEE.

[241] Honarvar, A. R. and Ghasem-Aghaee, N. (2009b). Casuist bdi-agent: a new extended bdi architecture with the capability of ethical reasoning. In *International Conference on Artificial Intelligence and Computational Intelligence*, pages 86–95, Berlin,Heidelberg. Springer.

[242] Honeycutt, D., Nourani, M., and Ragan, E. (2020). Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of

model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 63–72.

[243] Hong, J.-W. (2020). Why is artificial intelligence blamed more? analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction*, 36(18):1768–1774.

[244] Hong, J.-W. and Williams, D. (2019). Racism, responsibility and autonomy in hci: Testing perceptions of an ai agent. *Computers in Human Behavior*, 100:79–84.

[245] Honig, S. and Oron-Gilad, T. (2018). Understanding and resolving failures in human-robot interaction: Literature review and model development. *Frontiers in psychology*, 9:861.

[246] Horowitz, A. C. and Bekoff, M. (2007). Naturalizing anthropomorphism: Behavioral prompts to our humanizing of animals. *Anthrozoös*, 20(1):23–35.

[247] Horowitz, M. and Scharre, P. (2015). An introduction to autonomy in weapon systems. *Center for A New American Security Working Paper, February*.

[248] Horowitz, M. C. (2016). The ethics & morality of robotic warfare: Assessing the debate over autonomous weapons. *Daedalus*, 145(4):25–36.

[249] Hortensius, R., Hekele, F., and Cross, E. S. (2018). The perception of emotion in artificial agents. *IEEE Transactions on Cognitive and Developmental Systems*, 10(4):852–864.

[250] Horty, J. F. (2001). *Agency and deontic logic*. Oxford University Press, Oxford, UK.

[251] Horty, J. F. and Belnap, N. (1995). The deliberative stit: A study of action, omission, ability, and obligation. *Journal of philosophical logic*, 24(6):583–644.

[252] Hou, Y. T.-Y. and Jung, M. F. (2021). Who is the expert? reconciling algorithm aversion and algorithm appreciation in ai-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25.

[253] Howard, D. and Muntean, I. (2017). Artificial moral cognition: Moral functionalism and autonomous moral agency. In *Philosophy and Computing*, pages 121–159. Springer, Cham, Switzerland.

[254] Huang, X., Kwiatkowska, M., and Olejnik, M. (2019). Reasoning about cognitive trust in stochastic multiagent systems. *ACM Trans. Comput. Logic*, 20(4):21:1–21:64.

[255] Huang, X. and Kwiatkowska, M. Z. (2017). Reasoning about cognitive trust in stochastic multiagent systems. In Singh, S. P. and Markovitch, S., editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 3768–3774. AAAI Press.

[256] Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., and Garrabrant, S. (2019). Risks from learned optimization in advanced machine learning systems. *arXiv preprint arXiv:1906.01820*.

[257] Hübner, D. (2021). Two kinds of discrimination in ai-based penal decision-making. *ACM SIGKDD Explorations Newsletter*, 23(1):4–13.

[258] Hunter, P. (2019). The advent of ai and deep learning in diagnostics and imaging: Machine learning systems have potential to improve diagnostics in healthcare and imaging systems in research. *EMBO reports*, 20(7):e48559.

[259] Hursthouse, R. (1999). *On virtue ethics*. OUP Oxford.

[260] Hussain, R. and Zeadally, S. (2018). Autonomous cars: Research results, issues, and future challenges. *IEEE Communications Surveys & Tutorials*, 21(2):1275–1313.

[261] Hutchins, E. (2000). Distributed cognition. *International Encyclopedia of the Social and Behavioral Sciences. Elsevier Science*, 138.

[262] Hutchins, E. and Klausen, T. (1996). Distributed cognition in an airline cockpit. *Cognition and communication at work*, pages 15–34.

[263] Iklé, M., Franz, A., Rzepka, R., and Goertzel, B. (2018). *Artificial General Intelligence: 11th International Conference, AGI 2018, Prague, Czech Republic, August 22-25, 2018, Proceedings*. Springer, Heidelberg, Germany, 1 edition.

[264] Immanuel, K. (1785). Groundwork of the metaphysics of morals. In Radcliffe, E. S., McCarty, R., Allhoff, F., and Vaidya, A., editors, *Late Modern Philosophy: Essential Readings with Commentary*. Blackwell, New Jersey, USA.

[265] Inkpen, K. M. and Sedlins, M. (2011). Me and my avatar: exploring users' comfort with avatars for workplace communication. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work*, pages 383–386.

[266] Itti, L. and Borji, A. (2014). *Computational models: Bottom-up and top-down aspects.*, chapter 38. Oxford University Press.

[267] Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-ai symbiosis in organizational decision making. *Business Horizons*, 61(4):577–586.

[268] Javaid, M. and Estivill-Castro, V. (2021). Explanations from a robotic partner build trust on the robot's decisions for collaborative human-humanoid interaction. *Robotics*, 10(1).

[269] Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1):53–71.

[270] Jobin, A., Ienca, M., and Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399.

[271] Johnson, D. G. and Verdicchio, M. (2019). Ai, agency and responsibility: the vw fraud case and beyond. *AI & Society*, 34(3):639–647.

[272] Johnson, M. and Vera, A. (2019). No ai is an island: the case for teaming intelligence. *AI Magazine*, 40(1):16–28.

[273] Jung, M. F., Lee, J. J., DePalma, N., Adalgeirsson, S. O., Hinds, P. J., and Breazeal, C. (2013). Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, pages 1555–1566.

[274] Jussupow, E., Benbasat, I., and Heinzl, A. (2020). Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. *Research Papers*.

[275] Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134.

[276] Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., and Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, 125(2):131.

[277] Kamm, F. M. et al. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. OUP USA, Oxford, UK.

[278] Kant, I. (1785). *Groundwork of the metaphysics of morals*. Cambridge.

[279] Karacapilidis, N. I. and Pappis, C. P. (1997). A framework for group decision support systems: Combining ai tools and or techniques. *European Journal of Operational Research*, 103(2):373–388.

[280] Khademi, A. and Honavar, V. (2020). Algorithmic bias in recidivism prediction: A causal perspective (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13839–13840.

[281] Kim, P. H., Ferrin, D. L., Cooper, C. D., and Dirks, K. T. (2004). Removing the shadow of suspicion: the effects of apology versus denial for repairing competence-versus integrity-based trust violations. *Journal of applied psychology*, 89(1):104–118.

[282] Kim, T. and Song, H. (2021). How should intelligent agents apologize to restore trust? interaction effects between anthropomorphism and apology attribution on trust repair. *Telematics and Informatics*, 61:1–14.

[283] Klumpp, M. (2018). Automation and artificial intelligence in business logistics systems: human reactions and collaboration requirements. *International Journal of Logistics Research and Applications*, 21(3):224–242.

[284] Kneer, M. and Stuart, M. T. (2021). Playing the blame game with robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pages 407–411.

[285] Kocielnik, R., Amershi, S., and Bennett, P. N. (2019). *Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems*, page 1–14. Association for Computing Machinery, New York, NY, USA.

[286] Kohn, S. C., Momen, A., Wiese, E., Lee, Y.-C., and Shaw, T. H. (2019). The consequences of purposefulness and human-likeness on trust repair attempts made by self-driving vehicles. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 63, pages 222–226. SAGE Publications Sage CA: Los Angeles, CA.

[287] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86.

[288] Lahijanian, M. and Kwiatkowska, M. (2016). Social trust: a major challenge for the future of autonomous systems. In *2016 AAAI Fall Symposium Series*. AAAI Press.

[289] Langan-Fox, J., Canty, J. M., and Sankey, M. J. (2009). Human–automation teams and adaptable control for future air traffic management. *International Journal of Industrial Ergonomics*, 39(5):894–903.

[290] Larman, C. and Basili, V. R. (2003). Iterative and incremental developments. a brief history. *Computer*, 36(6):47–56.

[291] Law, J. (1992). Notes on the theory of the actor-network: Ordering, strategy, and heterogeneity. *Systems practice*, 5(4):379–393.

[292] Law, J. and Callon, M. (1988). Engineering and sociology in a military aircraft project: A network analysis of technological change. *Social problems*, 35(3):284–297.

[293] Lee, J. D. and See, K. A. (2004a). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1):50–80.

[294] Lee, J. D. and See, K. A. (2004b). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

[295] Lee, M. H., Siewiorek, D. P., Smailagic, A., Bernardino, A., and Bermúdez i Badia, S. (2021). A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

[296] Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., and Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 203–210, New York, NY,USA. IEEE.

[297] Letondal, C., Vinot, J.-L., Pauchet, S., Boussiron, C., Rey, S., Becquet, V., and Lavenir, C. (2018). Being in the sky: Framing tangible and embodied interaction for future airliner cockpits. In *Proceedings of the Twelfth International Conference on Tangible, Embedded, and Embodied Interaction*, pages 656–666.

[298] Lewandowsky, S., Mundy, M., and Tan, G. (2000). The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied*, 6(2):104.

[299] Lewicki, R. J. and Wiethoff, C. (2000). Trust, trust development, and trust repair. *The handbook of conflict resolution: Theory and practice*, 1(1):86–107.

[300] Lewis, M., Sycara, K., and Walker, P. (2018). The role of trust in human-robot interaction. In Abbass, H. A., Scholz, J., and Reid, D. J., editors, *Foundations of Trusted Autonomy*, pages 135–159. Springer International Publishing, Cham.

[301] Li, J., Zhao, X., Cho, M.-J., Ju, W., and Malle, B. F. (2016). From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. Technical report, SAE Technical Paper.

[302] Li, Z., Wang, Y., Wang, W., Greuter, S., and Mueller, F. F. (2020). Empowering a creative city: Engage citizens in creating street art through human-ai collaboration. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, page 1–8, New York, NY, USA. Association for Computing Machinery.

[303] Liao, Q. V., Wang, Y.-C., Bickmore, T., Fung, P., Grudin, J., Yu, Z., and Zhou, M. (2019). Human-agent communication: Connecting research and development in hci

and ai. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, pages 122–126.

[304] Lim, Y., Gardi, A., Sabatini, R., Ramasamy, S., Kistan, T., Ezer, N., Vince, J., and Bolia, R. (2018). Avionics human-machine interfaces and interactions for manned and unmanned aircraft. *Progress in Aerospace Sciences*, 102:1–46.

[305] Lima, G., Grgić-Hlača, N., and Cha, M. (2021a). Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–17.

[306] Lima, G., Grgic-Hlaca, N., and Cha, M. (2021b). Human perceptions on moral responsibility of AI: A case study in ai-assisted bail decision-making. *CoRR*, abs/2102.00625.

[307] Lindner, F., Bentzen, M. M., and Nebel, B. (2017). The hera approach to morally competent robots. In *IROS 2017: IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 6991–6997, Vancouver, BC, Canada. IEEE.

[308] Lindner, F., Mattmüller, R., and Nebel, B. (2019). Moral permissibility of action plans. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 7635–7642. AAAI Press.

[309] Liu, J., Gardi, A., Ramasamy, S., Lim, Y., and Sabatini, R. (2016). Cognitive pilot-aircraft interface for single-pilot operations. *Knowledge-based systems*, 112:37–53.

[310] Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103.

[311] Loi, M. and Christen, M. (2019). How to include ethics in machine learning research. *ERCIM News*, 116(3).

[312] Luo, X., Tong, S., Fang, Z., and Qu, Z. (2019). Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, 38(6):937–947.

[313] Lupton, D. (2016). *The quantified self: a sociology of sel-tracking*. Polity.

[314] Lyons, J. B. (2013). Being transparent about transparency: A model for human-robot interaction. In *2013 AAAI Spring Symposium Series*.

[315] Lyons, J. B. and Havig, P. R. (2014). Transparency in a human-machine context: Approaches for fostering shared awareness/intent. In Shumaker, R. and Lackey, S., editors, *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments*, pages 181–190, Cham. Springer International Publishing.

[316] Lyons, J. B., Sycara, K., Lewis, M., and Capiola, A. (2021). Human-autonomy teaming: Definitions, debates, and directions. *Frontiers in Psychology*, 12:1932.

[317] Ma, Y., Wang, Z., Yang, H., and Yang, L. (2020). Artificial intelligence applications in the development of autonomous vehicles: a survey. *IEEE/CAA Journal of Automatica Sinica*, 7(2):315–329.

[318] Machery, E. (2017). *Philosophy within its proper bounds*, chapter 2, pages 45–89. Oxford University Press, Oxford, UK.

[319] Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong*. Penguin Books, Frankfurt, Germany.

[320] Madl, T. and Franklin, S. (2015). Constrained incrementalist moral decision making for a biologically inspired cognitive architecture. In *A Construction Manual for Robots' Ethical Systems*, pages 137–153. Springer, Cham, Switzerland.

[321] Malatesta, L., Karpouzis, K., and Raouzaiou, A. (2009). Affective intelligence: the human face of ai. In *Artificial Intelligence An International Perspective*, pages 53–70. Springer.

[322] Malle, B. F. and Scheutz, M. (2019). Learning how to behave: Moral competence for social robots. *Handbuch Maschinenethik [Handbook of Machine Ethics]. Springer Reference Geisteswissenschaften. Wiesbaden, Germany: Springer. DOI*, 10:978–3.

[323] Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). Sacrifice one for the good of many?: People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pages 117–124. ACM.

[324] Malle, B. F., Scheutz, M., and Austerweil, J. L. (2017). Networks of social and moral norms in human and robot agents. In *A world with robots*, pages 3–17. Springer, Cham, Switzerland.

[325] Malle, B. F., Scheutz, M., Forlizzi, J., and Voiklis, J. (2016). Which robot am i thinking about?: The impact of action and appearance on people's evaluations of a moral robot. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, pages 125–132. IEEE Press.

[326] Malle, B. F. and Ullman, D. (2020). A multi-dimensional conception and measure of human-robot trust.

[327] Malle, B. F. and Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction*, pages 3–25. Elsevier.

[328] Marchant, G. E. (2017). Artificial Intelligence and the Future of Legal Practice. *Scitech Lawyer*, 14(1):20–23. Num Pages: 20-23 Place: Chicago, United States Publisher: American Bar Association.

[329] Marinaccio, K., Kohn, S., Parasuraman, R., and De Visser, E. J. (2015). A framework for rebuilding trust in social automation across health-care domains. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, volume 4, pages 201–205, New Delhi, India. SAGE Publications Sage India.

[330] Markowsky, J. K. (1975). Why anthropomorphism in children's literature? *Elementary English*, 52(4):460–466.

[331] Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology*, 6(3):175–183.

[332] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3):709.

[333] McGloin, R., Nowak, K. L., Stiffano, S. C., and Flynn, G. M. (2009). The effect of avatar perception on attributions of source and text credibility. In *Proceedings of ISPR*

2009 The International Society for Presence Research Annual Conference. Philadelphia: Temple University Press.

[334] McGregor, M. and Tang, J. C. (2017). More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, pages 2208–2220, New York, NY, USA. Association for Computing Machinery.

[335] McKnight, D. H. and Chervany, N. L. (2001). Trust and distrust definitions: One bite at a time. In *Trust in Cyber-societies*, pages 27–54. Springer.

[336] McLaren, B. M. (2003). Extensionally defining principles and cases in ethics: An ai model. *Artificial Intelligence*, 150(1-2):145–181.

[337] McNeese, N. J., Demir, M., Cooke, N. J., and Myers, C. (2018). Teaming with a synthetic teammate: Insights into human-autonomy teaming. *Human factors*, 60(2):262–273.

[338] McNeese, N. J., Demir, M., Cooke, N. J., and She, M. (2021). Team Situation Awareness and Conflict: A Study of Human–Machine Teaming. *Journal of Cognitive Engineering and Decision Making*.

[339] Meertens, R. M. and Lion, R. (2008). Measuring an individual's tendency to take risks: the risk propensity scale 1. *Journal of Applied Social Psychology*, 38(6):1506–1520.

[340] Mermet, B. and Simon, G. (2016). Formal verication of ethical properties in multiagent systems. In *1st Workshop on Ethics in the Design of Intelligent Agents*, pages 26–31, The Hague, Netherlands. CEUR.

[341] Merrill, N. and Cheshire, C. (2017). Trust your heart: Assessing cooperation and trust with biosignals in computer-mediated interactions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 2–12.

[342] Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in cognitive sciences*, 11(4):143–152.

[343] Miller, A. (2003). *An Introduction to Contemporary Metaethics*. Polity, Cambridge, UK.

[344] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

[345] Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., and Tscheligi, M. (2017). To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI*, 4:21.

[346] Mitchell, R. W., Thompson, N. S., and Miles, H. L., editors (1997). *Anthropomorphism, anecdotes, and animals*. Suny Press.

[347] Monti, M. (2019). Automated journalism and freedom of information: Ethical and juridical problems related to ai in the press field. *Opinio Juris in Comparatione*, 1:2018.

[348] Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems*, 21(4):18–21.

[349] Mosier, K. L., Fischer, U., Burian, B. K., and Kochan, J. A. (2017). Autonomous, context-sensitive, task management systems and decision support tools i: Human-autonomy teaming fundamentals and state of the art. Technical report, NASA.

[350] Mukherjee, R., Jonsdottir, G., Sen, S., and Sarathi, P. (2001). Movies2go: an online voting based movie recommender system. In *Proceedings of the fifth international conference on Autonomous agents*, pages 114–115.

[351] Mumaw, R. J. and Holder, B. E. (2002). What do cultural dimensions reveal about flight deck operations? In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 46, pages 230–234. SAGE Publications Sage CA: Los Angeles, CA.

[352] Murakami, Y. (2005). Utilitarian deontic logic. *Advances in Modal Logic*, 5:211–230.

[353] Murray, S. L., Holmes, J. G., and Collins, N. L. (2006). Optimizing assurance: The risk regulation system in relationships. *Psychological bulletin*, 132(5):641.

[354] Mynatt, C. and Sherman, S. J. (1975). Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology*, 32(6):1111.

[355] Nahavandi, S. (2017). Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. *IEEE Systems, Man, and Cybernetics Magazine*, 3(1):10–17.

[356] Nasirian, F., Ahmadian, M., and Lee, O.-K. D. (2017). Ai-based voice assistant systems: Evaluating from the interaction and trust perspectives. In *AMCIS 2017 Proceedings*.

[357] Nass, C., Steuer, J., and Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78.

[358] Neff, G. and Nafus, D. (2016). *Self-tracking*. The MIT Press essential knowledge series. The MIT Press.

[359] Neto, B. F. d. S., da Silva, V. T., and de Lucena, C. J. P. (2011). Nbdi: An architecture for goal-oriented normative agents. In *ICAART 2011 - Proceedings of the 3rd International Conference on Agents and Artificial Intelligence*, volume 1, pages 116–125, Berlin,Heidelberg. Springer.

[360] Newman, D. T., Fast, N., and Harmon, D. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, 160:149–167.

[361] Niforatos, E., Palma, A., Gluszny, R., Vourvopoulos, A., and Liarokapis, F. (2020). *Would You Do It?: Enacting Moral Dilemmas in Virtual Reality for Understanding Ethical Decision-Making*, page 1–12. Association for Computing Machinery, New York, NY, USA.

[362] Nomura, S., Hutchins, E., and Holder, B. E. (2006). The uses of paper in commercial airline flight operations. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, pages 249–258.

[363] Noothigattu, R., Gaikwad, S., Awad, E., Dsouza, S., Rahwan, I., Ravikumar, P., and Procaccia, A. D. (2018). A voting-based system for ethical decision making. In *Proceedings of the thirty-second AAAI conference on artificial intelligence 2018*, pages 1587–1594, Palo Alto, California, USA. AAAI Press.

[364] Nordqvist, M. and Lindblom, J. (2018). Operators' experience of trust in manual assembly with a collaborative robot. In *Proceedings of the 6th International Conference on Human-Agent Interaction*, pages 341–343, New York, NY, USA. ACM, ACM.

[365] Nourani, M., Kabir, S., Mohseni, S., and Ragan, E. D. (2019). The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 97–105.

[366] Nourani, M., King, J., and Ragan, E. (2020). The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 112–121.

[367] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al. (2020). Bias in data-driven artificial intelligence systems—an introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(3):e1356.

[368] of the Red Cross, I. R. (2019). Artificial intelligence and machine learning in armed conflict: A human-centred approach. Technical Report 102, ICRC.

[369] Osmani, N. et al. (2020). The complexity of criminal liability of ai systems. *Masaryk University Journal of Law and Technology*, 14(1):53–82.

[370] Ozgur, A. (2004). Supervised and unsupervised machine learning techniques for text document categorization. *Unpublished Master's Thesis, İstanbul: Boğaziçi University.*

[371] O'Neill, T., McNeese, N., Barron, A., and Schelble, B. (2020). Human–Autonomy Teaming: A Review and Analysis of the Empirical Literature. *Human Factors*, pages 1–35.

[372] Paliouras, G., Papatheodorou, C., Karkaletsis, V., and Spyropoulos, C. D. (2002). Discovering user communities on the internet using unsupervised machine learning techniques. *Interacting with Computers*, 14(6):761–791.

[373] Panganiban, A. R., Matthews, G., and Long, M. D. (2020). Transparency in Autonomous Teammates: Intention to Support as Teaming Information. *Journal of Cognitive Engineering and Decision Making*, 14(2):174–190.

[374] Papenmeier, A., Englebienne, G., and Seifert, C. (2019). How model accuracy and explanation fidelity influence user trust in ai. In *IJCAI Workshop on Explainable Artificial Intelligence (XAI) 2019.*

[375] Paraiso, E. C. and Tacla, C. A. (2009). Using embodied conversational assistants to interface users with multi-agent based cscw applications: The webanima agent. *J. Univers. Comput. Sci.*, 15(9):1991–2010.

[376] Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 30(3):286–297.

[377] Park, H., Ahn, D., Hosanagar, K., and Lee, J. (2021). Human-ai interaction in human resource management: Understanding why employees resist algorithmic evaluation at workplaces and how to mitigate burdens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

[378] Parthemore, J. and Whitby, B. (2014). Moral agency, moral responsibility, and artifacts: What existing artifacts fail to achieve (and why), and why they, nevertheless, can (and do!) make moral claims upon us. *International Journal of Machine Consciousness*, 6(02):141–161.

[379] Pauchet, S., Vinot, J.-L., Letondal, C., Lemort, A., Lavenir, C., Lecomte, T., Rey, S., Becquet, V., and Crouzet, G. (2019). Multi-plié: A linear foldable and flattenable interactive display to support efficiency, safety and collaboration. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13. ACM.

[380] Pereira, L. M. and Saptawijaya, A. (2007). Modelling morality with prospective logic. In *Portuguese Conference on Artificial Intelligence*, pages 99–111, Berlin,Heidelberg. Springer.

[381] Pereira, L. M. and Saptawijaya, A. (2011). *Modelling morality with prospective logic*, pages 398–421. Cambridge University Press, Cambridge,UK.

[382] Pereira, L. M. and Saptawijaya, A. (2016). *Programming machine ethics*, volume 26. Springer, Cham, Switzerland.

[383] Pereira, L. M. and Saptawijaya, A. (2017). Counterfactuals, logic programming and agent morality. In *Applications of Formal Philosophy*, pages 25–53. Springer, Berlin, Heidelberg.

[384] Phillips, E., Schaefer, K. E., Billings, D. R., Jentsch, F., and Hancock, P. A. (2016). Human-animal teams as an analog for future human-robot teams: Influencing design and fostering trust. *Journal of Human-Robot Interaction*, 5(1):100–125.

[385] Piano, S. L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*, 7(1):1–7.

[386] Picard, R. W. (2000). *Affective computing*. MIT press.

[387] Picard, R. W. et al. (1995). *Affective computing*. MIT Press, Cambridge, MA, USA.

[388] Pini, A., Hayes, J., Upton, C., and Corcoran, M. (2019). Ai inspired recipes: Designing computationally creative food combos. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, page 1–6, New York, NY, USA. Association for Computing Machinery.

[389] Pink, S., Ruckenstein, M., Willim, R., and Duque, M. (2018). Broken data: Conceptualising data in an emerging world. *Big Data & Society*, 5(1):1–13.

[390] Pinto, J. and Reiter, R. (1993). Temporal reasoning in logic programming: A case for the situation calculus. In *Proceedings of the 10th International Conference in Logic Programming*, volume 93, pages 203–221, Berlin,Heidelberg. Springer.

[391] Planke, L. J., Lim, Y., Gardi, A., Sabatini, R., Kistan, T., and Ezer, N. (2020). A cyber-physical-human system for one-to-many uas operations: Cognitive load analysis. *Sensors*, 20(19):5467.

[392] Poltrock, S., Handel, M. J., Poteet, S. R., and Murray, P. (2012). Recognizing team context during simulated missions. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 197–206.

[393] Pontier, M. and Hoorn, J. (2012). Toward machines that behave ethically better than humans do. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, pages 2198–2203, Austin, TX:. Cognitive Science Society.

[394] Pontier, M. A. and Widdershoven, G. A. M. (2013). Robots that stimulate autonomy. In Papadopoulos, H., Andreou, A. S., Iliadis, L., and Maglogiannis, I., editors, *Artificial Intelligence Applications and Innovations*, pages 195–204, Berlin, Heidelberg. Springer Berlin Heidelberg.

[395] Porcheron, M., Clark, L., Jones, M., Candello, H., Cowan, B. R., Murad, C., Sin, J., Aylett, M. P., Lee, M., Munteanu, C., et al. (2020). Cui@ cscw: Collaborating through conversational user interfaces. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing*, pages 483–492.

[396] Porcheron, M., Fischer, J. E., McGregor, M., Brown, B., Luger, E., Candello, H., and O'Hara, K. (2017a). Talking with conversational agents in collaborative action. In *Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 431–436.

[397] Porcheron, M., Fischer, J. E., and Sharples, S. (2017b). " do animals have accents?" talking with agents in multi-party conversation. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 207–219.

[398] Poria, S., Cambria, E., Bajpai, R., and Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125.

[399] Poushneh, A. (2020). Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services*, 58:102283.

[400] Price, I. and Nicholson, W. (2019). Medical ai and contextual bias. *Harv. J.L.*

[401] Prinz, J. (2007). *The Emotional Construction of Morals*. Oxford University Press, Oxford, UK.

[402] Promberger, M. and Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5):455–468.

[403] Qiu, L. and Benbasat, I. (2005). Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International journal of human-computer interaction*, 19(1):75–94.

[404] Rachels, J. (1979). *Active and passive euthanasia*, pages 551–516. Springer, Boston, MA, USA.

[405] Ragot, M., Martin, N., and Cojean, S. (2020). Ai-generated vs. human artworks. a perception bias towards artificial intelligence? In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*, pages 1–10.

[406] Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J.-F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., et al. (2019). Machine behaviour. *Nature*, 568(7753):477–486.

[407] Rao, Q. and Frtunikj, J. (2018). Deep learning for self-driving cars: Chances and challenges. In *Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems*, pages 35–38.

[408] Rau, P. P., Li, Y., and Li, D. (2009). Effects of communication style and culture on ability to accept recommendations from robots. *Computers in Human Behavior*, 25(2):587–595.

[409] Rawls, J. (2009). *A theory of justice*. Harvard University Press, London, UK.

[410] Reason, J. (1990). *Human error*. Cambridge university press.

[411] Reed, G. S., Petty, M. D., Jones, N. J., Morris, A. W., Ballenger, J. P., and Delugach, H. S. (2016). A principles-based model of ethical considerations in military decision making. *The Journal of Defense Modeling and Simulation*, 13(2):195–211.

[412] Rempel, J. K., Holmes, J. G., and Zanna, M. P. (1985). Trust in close relationships. *Journal of personality and social psychology*, 49(1):95.

[413] Richards, D. and Amos, M. (2014). Evolving morphologies with cppn-neat and a dynamic substrate. In *Artificial Life Conference Proceedings 14*, pages 255–262, address=Cambridge, MA, USA. MIT Press.

[414] Richards, R. A. (2002). Application of multiple artificial intelligence techniques for an aircraft carrier landing decision support tool. In *Proceedings of 2002 IEEE International Conference on Fuzzy Systems*, volume 1, pages 7–11. IEEE.

[415] Ritov, I. and Baron, J. (1992). Status-quo and omission biases. *Journal of Risk and Uncertainty*, 5:49–61.

[416] Robinette, P., Howard, A. M., and Wagner, A. R. (2015). Timing is key for robot trust repair. In *International Conference on Social Robotics*, pages 574–583. Springer.

[417] Robinette, P., Li, W., Allen, R., Howard, A. M., and Wagner, A. R. (2016). Overtrust of robots in emergency evacuation scenarios. In *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, pages 101–108, Piscataway, NJ, USA. IEEE Press.

[418] Rowe, W. D. (1975). *An" Anatomy" of risk*. Environmental Protection Agency.

[419] Roy, N., Torre, M. V., Gadiraju, U., Maxwell, D., and Hauff, C. (2021). Note the highlight: Incorporating active reading tools in a search as learning environment. In

Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, pages 229–238.

[420] Rusbult, C. E., Verette, J., Whitney, G. A., Slovik, L. F., and Lipkus, I. (1991). Accommodation processes in close relationships: Theory and preliminary empirical evidence. *Journal of Personality and social Psychology*, 60(1):53.

[421] Russell, S. and Norvig, P. (2010). *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3 edition.

[422] Rzepka, R. and Araki, K. (2017). What people say? web-based casuistry for artificial morality experiments. In *International Conference on Artificial General Intelligence*, pages 178–187, Cham, Switzerland. Springer.

[423] Saad, L. (2010). Four moral issues sharply divide americans. https://news.gallup.com/poll/137357/four-moral-issues-sharply-divide-americans.aspx. [Online; accessed 06-December-2019].

[424] Šabanović, S., Bennett, C. C., Chang, W.-L., and Huber, L. (2013). Paro robot affects diverse interaction modalities in group sensory therapy for older adults with dementia. In *2013 IEEE 13th international conference on rehabilitation robotics (ICORR)*, pages 1–6. IEEE.

[425] Sachdeva, S., Singh, P., and Medin, D. (2011). Culture and the quest for universal principles in moral reasoning. *International journal of psychology*, 46(3):161–176.

[426] Salas, E., Prince, C., Baker, D. P., and Shrestha, L. (1995). Situation Awareness in Team Performance: Implications for Measurement and Training. *Human Factors*, 37(1):123–136.

[427] Salles, A., Evers, K., and Farisco, M. (2020). Anthropomorphism in AI. *AJOB neuroscience*, 11(2):88–95.

[428] Sanders, D. and Gegov, A. (2013). Ai tools for use in assembly automation and some examples of recent applications. *Assembly Automation*.

[429] Sannon, S., Stoll, B., DiFranzo, D., Jung, M., and Bazarova, N. N. (2018). How personification and interactivity influence stress-related disclosures to conversational agents. In *companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 285–288.

[430] Santoni de Sio, F. and Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, 5:15.

[431] Saptawijaya, A. and Pereira, L. M. (2014). Towards modeling morality computationally with logic programming. In *International Symposium on Practical Aspects of Declarative Languages*, pages 104–119, Berlin,Heidelberg. Springer.

[432] Saptawijaya, A. and Pereira, L. M. (2015). The potential of logic programming as a computational tool to model morality. In *A Construction Manual for Robots' Ethical Systems*, pages 169–210. Springer, Berlin,Heidelberg.

[433] Saptawijaya, A., Pereira, L. M., et al. (2012). Moral reasoning under uncertainty. In *International Conference on Logic for Programming Artificial Intelligence and Reasoning*, pages 212–227, Berlin,Heidelberg. Springer.

[434] Sari, R. F., Rochim, A. F., Tangkudung, E., Tan, A., and Marciano, T. (2017). Location-based mobile application software development: Review of waze and other apps. *Advanced Science Letters*, 23(3):2028–2032.

[435] Satapathy, S. M., Jhaveri, R., Khanna, U., and Dwivedi, A. K. (2020). Smart rent portal using recommendation system visualized by augmented reality. *Procedia Computer Science*, 171:197–206.

[436] Sauppé, A. (2014). Designing effective strategies for human-robot collaboration. In *Proceedings of the Companion Publication of the 17th ACM Conference on Computer Supported Cooperative Work & social computing*, pages 85–88.

[437] Schaefer, K. E., Chen, J. Y., Szalma, J. L., and Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors*, 58(3):377–400.

[438] Scheeff, M., Pinto, J., Rahardja, K., Snibbe, S., and Tow, R. (2002). Experiences with sparky, a social robot. In *Socially intelligent agents*, pages 173–180. Springer.

[439] Schelble, B. G., Flathmann, C., and McNeese, N. (2020). Towards Meaningfully Integrating Human-Autonomy Teaming in Applied Settings. In *Proceedings of the 8th International Conference on Human-Agent Interaction*, HAI '20, pages 149–156, New York, NY, USA. Association for Computing Machinery.

[440] Scheutz, M. and Malle, B. F. (2014). Think and do the right thing: a plea for morally competent autonomous robots. In *Proceedings of the IEEE 2014 International Symposium on Ethics in Engineering, Science, and Technology*, page 9, New York, NY, USA. IEEE Press.

[441] Schmitt, A., Wambsganss, T., Söllner, M., and Janson, A. (2021). Towards a trust reliance paradox? exploring the gap between perceived trust in and reliance on algorithmic advice. In *International Conference on Information Systems (ICIS)*.

[442] Schouten, J. W. and McAlexander, J. H. (1995). Subcultures of consumption: An ethnography of the new bikers. *Journal of consumer research*, 22(1):43–61.

[443] Schwaninger, I., Fitzpatrick, G., and Weiss, A. (2019). Exploring trust in human-agent collaboration. In *Proceedings of 17th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).

[444] Schweitzer, M. E., Hershey, J. C., and Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, 101(1):1–19.

[445] Schwitzgebel, E. and Cushman, F. (2012). Expertise in moral reasoning? order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language*, 27(2):135–153.

[446] Schwitzgebel, E. and Cushman, F. (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition*, 141:127–137.

[447] Sebo, S. S., Krishnamurthi, P., and Scassellati, B. (2019). "i don't believe you": Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 57–65. IEEE.

[448] Sengupta, S., Donekal, A. K., and Mathur, A. R. (2016). Automation in modern airplanes-a safety and human factors based study. In *INCOSE International Symposium*, volume 26, pages 386–394. Wiley Online Library.

[449] Serpell, J. (2003). Anthropomorphism and anthropomorphic selection—beyond the" cute response". *Society & Animals*, 11(1):83–100.

[450] Shanmuganathan, M. (2020). Behavioural finance in an era of artificial intelligence: Longitudinal case study of robo-advisors in investment decisions. *Journal of Behavioral and Experimental Finance*, 27:100297.

[451] Shappell, S. A. and Wiegmann, D. A. (2000). The human factors analysis and classification system–hfacs. Technical report, FAA Civil Aeromedical Institute and University of Illinois at Urbana-Champaign, Institute of Aviation.

[452] SharifHeravi, M., Taylor, J. R., Stanton, C. J., Lambeth, S., and Shanahan, C. (2020). It'sa disaster! factors affecting trust development and repair following agent task failure. In *Proceedings of the 2020 Australasian Conference on Robotics and Automation (ACRA 2020), 8-10 December 2020, Brisbane, Queensland.*

[453] Sheller, M. (2004). Automotive emotions: Feeling the car. *Theory, Culture & Society*, 21(4-5):221–242.

[454] Sheridan, T. B. and Verplank, W. L. (1978). Human and Computer Control of Undersea Teleoperators. Technical report, Massachusetts inst of tech Cambridge man-machine systems lab.

[455] Shim, J., Arkin, R., and Pettinatti, M. (2017). An intervening ethical governor for a robot mediator in patient-caregiver relationship: Implementation and evaluation. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on Robotics and Automation*, pages 2936–2942, New York, NY, USA. IEEE.

[456] Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies*, 146:102551.

[457] Shively, R. J., Lachter, J., Brandt, S. L., Matessa, M., Battiste, V., and Johnson, W. W. (2018a). Why Human-Autonomy Teaming? In Baldwin, C., editor, *Advances in Neuroergonomics and Cognitive Engineering*, Advances in Intelligent Systems and Computing, pages 3–11, Cham. Springer International Publishing.

[458] Shively, R. J., Lachter, J., Koteskey, R., and Brandt, S. L. (2018b). Crew Resource Management for Automated Teammates (CRM-A). In Harris, D., editor, *Engineering Psychology and Cognitive Ergonomics*, Lecture Notes in Computer Science, pages 215–229, Cham. Springer International Publishing.

[459] Shortliffe, E. H. (1977). Mycin: A knowledge-based computer program applied to infectious diseases. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 66. American Medical Informatics Association.

[460] Siau, K. and Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal*, 31(2):47–53.

[461] Sinnott-Armstrong, W. (2021). Consequentialism. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition.

[462] Skidmore, M., Zenyuh, J., Small, R., Quinn, T., and Moyers, C. (1995). Dual use applications of hazard monitoring: commercial and military aviation and beyond. In *Proceedings of the IEEE 1995 National Aerospace and Electronics Conference. NAECON 1995*, volume 2, pages 948–955. IEEE.

[463] Sparrow, R. (2007). Killer robots. *Journal of applied philosophy*, 24(1):62–77.

[464] Spencer, J., Poggi, J., and Gheerawo, R. (2018). Designing out stereotypes in artificial intelligence: Involving users in the personality design of a digital assistant. In *Proceedings of the 4th EAI International Conference on Smart Objects and Technologies for Social Good*, Goodtechs '18, page 130–135, New York, NY, USA. Association for Computing Machinery.

[465] Spranca, M., Minsk, E., and Baron, J. (1991). Omission and commission in judgment and choice. *Journal of Experimental Social Psychology*, 27(1):76–105.

[466] Sreenivasan, G. (2002). Errors about errors: Virtue theory and trait attribution. *Mind*, 111(441):47–68.

[467] Stahl, B. C. and Wright, D. (2018). Ethics and privacy in ai and big data: Implementing responsible research and innovation. *IEEE Security & Privacy*, 16(3):26–33.

[468] Stanley, K. O., D'Ambrosio, D. B., and Gauci, J. (2009). A hypercube-based encoding for evolving large-scale neural networks. *Artificial life*, 15(2):185–212.

[469] Stanley, K. O. and Miikkulainen, R. (2002). Evolving neural networks through augmenting topologies. *Evolutionary computation*, 10(2):99–127.

[470] Stanton, N. A. and Salmon, P. M. (2009). Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems. *Safety Science*, 47(2):227–237.

[471] Stevens, C. J., Pinchbeck, B., Lewis, T., Luerssen, M., Pfitzner, D., Powers, D. M., Abrahamyan, A., Leung, Y., and Gibert, G. (2016). Mimicry and expressiveness of an eca in human-agent interaction: familiarity breeds content! *Computational cognitive science*, 2(1):1–14.

[472] Stuart, M. T. and Kneer, M. (2021). Guilty artificial minds: Folk attributions of mens rea and culpability to artificially intelligent agents. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–27.

[473] Sullins, J. P. (2006). When is a robot a moral agent? *IRIE*, 6:23–30.

[474] Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative.

[475] Surber, R. (2018). Artificial intelligence: Autonomous technology (at), lethal autonomous weapons systems (laws) and peace time threats. *ICT4Peace Foundation and the Zurich Hub for Ethics and Technology (ZHET) p*, 1:21.

[476] Susskind, A. M. and Odom-Reed, P. R. (2019). Team Member's Centrality, Cohesion, Conflict, and Performance in Multi-University Geographically Distributed Project Teams. *Communication Research*, 46(2):151–178.

[477] Suwandana, I. G. M. (2019). Role of transformational leadership mediation: Effect of emotional and communication intelligence towards teamwork effectiveness. *International research journal of management, IT and social sciences*, 6(2):52–62.

[478] Taddeo, M. and Floridi, L. (2018). How ai can be a force for good. *Science*, 361(6404):751–752.

[479] Thagard, P. (2019). *Mind-Society: From Brains to Social Sciences and Professions (Treatise on Mind and Society)*. Oxford University Press.

[480] Theodoridis, T., Solachidis, V., Dimitropoulos, K., Gymnopoulos, L., and Daras, P. (2019). A survey on ai nutrition recommender systems. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 540–546.

[481] Thimbleby, H., Marsh, S., Jones, S., and Cockburn, A. (2018). Trust in cscw. In *Computer-Supported Cooperative Work*, pages 253–271. Routledge.

[482] Thomas, L. (1987). Friendship. *Synthese*, 72(2):217–236.

[483] Thomson, A. L. (2017). Investigating trust and trust recovery in human-robot interactions. In *Augustana Celebration of Learning*.

[484] Thornton, S. M., Pan, S., Erlien, S. M., and Gerdes, J. C. (2017). Incorporating ethical considerations into automated vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 18(6):1429–1439.

[485] Thurman, N., Moeller, J., Helberger, N., and Trilling, D. (2019). My friends, editors, algorithms, and i: Examining audience attitudes to news selection. *Digital Journalism*, 7(4):447–469.

[486] Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, 34(3):589–607.

[487] Toader, D.-C., Boca, G., Toader, R., Măcelaru, M., Toader, C., Ighian, D., Rădulescu, A. T., et al. (2019). The effect of social presence and chatbot errors on trust. *Sustainability*, 12(1):1–1.

[488] Tokadlı, G., Dorneich, M. C., and Matessa, M. (2021). Evaluation of playbook delegation approach in human-autonomy teaming for single pilot operations. *International Journal of Human–Computer Interaction*, 37:703–716.

[489] Tokushige, H., Narumi, T., Ono, S., Fuwamoto, Y., Tanikawa, T., and Hirose, M. (2017). Trust lengthens decision time on unexpected recommendations in human-agent interaction. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 245–252.

[490] Tolmeijer, S., Gadiraju, U., Ghantasala, R., Gupta, A., and Bernstein, A. (2021a). Second chance for a first impression? trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*, pages 77–87.

[491] Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., and Bernstein, A. (2020a). Implementations in machine ethics: a survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38.

[492] Tolmeijer, S., Weiss, A., Hanheide, M., Lindner, F., Powers, T. M., Dixon, C., and Tielman, M. L. (2020b). Taxonomy of trust-relevant failures and mitigation strategies. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 3–12.

[493] Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M., and Bernstein, A. (2021b). Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

[494] Tuckman, B. W. (1965). Developmental Sequence in Small Groups. *Psychological Bulletin*, 63(6):384–399.

[495] Tufiş, M. and Ganascia, J.-G. (2015). Grafting norms onto the bdi agent model. In *A Construction Manual for Robots' Ethical Systems*, pages 119–133. Springer, Cham, Switzerland.

[496] Turilli, M. (2007). Ethical protocols design. *Ethics and Information Technology*, 9(1):49–62.

[497] Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *science*, 185(4157):1124–1131.

[498] Tversky, A. and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323.

[499] Umbrello, S., Torres, P., and De Bellis, A. F. (2020). The future of war: could lethal autonomous weapons make conflict more ethical? *AI & Society*, 35(1):273–282.

[500] Van Dang, C., Tran, T. T., Gil, K.-J., Shin, Y.-B., Choi, J.-W., Park, G.-S., and Kim, J.-W. (2017). Application of soar cognitive agent based on utilitarian ethics theory for home service robots. In *Ubiquitous Robots and Ambient Intelligence (URAI), 2017 14th International Conference on*, pages 155–158, New York, NY, USA. IEEE.

[501] Van de Voort, M., Pieters, W., and Consoli, L. (2015). Refining the ethics of computer-made decisions: a classification of moral mediation by ubiquitous machines. *Ethics and Information Technology*, 17(1):41–56.

[502] Vanderelst, D. and Winfield, A. (2017). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research*, 48:56–66.

[503] Velleman, J. D. (2013). *Foundations for Moral Relativism.* OpenBook Publishers, Cambridge, UK.

[504] Verger, R. (2019). Explore the gauges, levers, and history of a 747's iconic cockpit. Populare Science, published online.

[505] Verheij, B. (2016). Formalizing value-guided argumentation for ethical systems design. *Artificial Intelligence and Law*, 24(4):387–407.

[506] Vodrahalli, K., Gerstenberg, T., and Zou, J. (2021). Do humans trust advice more if it comes from ai? an analysis of human-ai interactions. *CoRR*, abs/2107.07015.

[507] Voiklis, J., Kim, B., Cusimano, C., and Malle, B. F. (2016). Moral judgments of human vs. robot agents. In *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 775–780, New York, NY, USA. IEEE.

[508] von der Lieth Gardner, A. (1987). *An artificial intelligence approach to legal reasoning*. The MIT Press.

[509] Wagner, A. R., Robinette, P., and Howard, A. (2018). Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Trans. Interact. Intell. Syst.*, 8(4):26:1–26:24.

[510] Wallach, M. A., Kogan, N., and Bem, D. J. (1964). Diffusion of responsibility and level of risk taking in groups. *The Journal of Abnormal and Social Psychology*, 68(3):263.

[511] Wallach, W. (2010). Cognitive models of moral decision making. *Topics in cognitive science*, 2(3):420–429.

[512] Wallach, W. (2017). Toward a ban on lethal autonomous weapons: surmounting the obstacles. *Communications of the ACM*, 60(5):28–34.

[513] Wallach, W. and Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. Oxford University Press, Oxford, UK.

[514] Wallach, W., Allen, C., and Smit, I. (2008). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & Society*, 22(4):565–582.

[515] Wallach, W., Franklin, S., and Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in cognitive science*, 2(3):454–485.

[516] Wang, N., Pynadath, D. V., and Hill, S. G. (2016). Trust calibration within a human-robot team: Comparing automatically generated explanations. In *11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 109–116, New York, NY,USA. IEEE.

[517] Wang, Y.-H., Li, Y., Yang, S.-L., and Yang, L. (2005). Classification of substrates and inhibitors of p-glycoprotein using unsupervised machine learning approach. *Journal of chemical information and modeling*, 45(3):750–757.

[518] Watson, D. (2019). The rhetoric and reality of anthropomorphism in artificial intelligence. *Minds and Machines*, 29(3):417–440.

[519] Webster, M., Western, D., Araiza-Illan, D., Dixon, C., Eder, K., Fisher, M., and Pipe, A. (2020). A corroborative approach to verification and validation of human–robot teams. *International Journal of Robotics Research*, 39:73–99.

[520] Wellsandta, S., Rusak, Z., Ruiz Arenas, S., Aschenbrenner, D., Hribernik, K. A., and Thoben, K.-D. (2020). Concept of a Voice-Enabled Digital Assistant for Predictive Maintenance in Manufacturing. SSRN Scholarly Paper ID 3718008, Social Science Research Network, Rochester, NY.

[521] Whitley, R. (2000). *The intellectual and social publisher of the sciences*. Oxford University Press on Demand, Oxford, UK.

[522] Wiegel, V. and van den Berg, J. (2009). Combining moral theory, modal logic and mas to create well-behaving artificial agents. *International Journal of Social Robotics*, 1(3):233–242.

[523] Wiltshire, T. J., Warta, S. F., Barber, D., and Fiore, S. M. (2017). Enabling robotic social intelligence by engineering human social-cognitive mechanisms. *Cognitive Systems Research*, 43:190–207.

[524] Winfield, A. F., Blum, C., and Liu, W. (2014). Towards an ethical robot: internal models, consequences and ethical action selection. In *Conference Towards Autonomous Robotic Systems*, pages 85–96, Cham, Switzerland. Springer.

[525] Wolkenstein, A. (2018). What has the trolley dilemma ever done for us (and what will it do in the future)? on some recent debates about the ethics of self-driving cars. *Ethics and Information Technology*, 20(3):163–173.

[526] Wölker, A. and Powell, T. E. (2021). Algorithms in the newsroom? news readers' perceived credibility and selection of automated journalism. *Journalism*, 22(1):86–103.

[527] Wollert, M. (2018). Public perception of autonomous aircraft. *Order*, 10810632.

[528] Wong, S., Baltuch, G., Jaggi, J., and Danish, S. (2009). Functional localization and visualization of the subthalamic nucleus from microelectrode recordings acquired during dbs surgery with unsupervised machine learning. *Journal of neural engineering*, 6(2):1–11.

[529] Wooldridge, M. and Van Der Hoek, W. (2005). On obligations and normative ability: Towards a logical analysis of the social contract. *Journal of Applied Logic*, 3(3-4):396–420.

[530] Woollard, F. (2012). The doctrine of doing and allowing ii: The moral relevance of the doing/allowing distinction. *Philosophy Compass*, 7(7):459–469.

[531] Woollard, F. (2015). *Doing and allowing harm*. Oxford University Press, USA, Kettering, Northants, USA.

[532] Wu, Y.-H. and Lin, S.-D. (2017). A low-cost ethics shaping approach for designing reinforcement learning agents. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 1687–1694, Palo Alto, California, USA. AAAI Press.

[533] Wynne, C. D. (2004). The perils of anthropomorphism. *Nature*, 428(6983):606–606.

[534] Wynne, K. T. and Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, 19(3):353–374.

[535] Xu, L., Zhou, X., and Gadiraju, U. (2020). How does team composition affect knowledge gain of users in collaborative web search? In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, pages 91–100.

[536] Xu, S., Tan, W., Sun, L., and Qu, X. (2017). Survey on theory and method of pilot-aircraft system with intelligent control. In *2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE)*, pages 92–96. IEEE.

[537] Yang, Q., Steinfeld, A., Rosé, C., and Zimmerman, J. (2020). *Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design*, page 1–13. Association for Computing Machinery, New York, NY, USA.

[538] Yang, Q., Steinfeld, A., and Zimmerman, J. (2019). *Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes*, page 1–11. Association for Computing Machinery, New York, NY, USA.

[539] Yarger, L., Payton, F. C., and Neupane, B. (2019). Algorithmic equity in the hiring of underrepresented it job candidates. *Online Information Review*.

[540] Yilmaz, L., Franco-Watkins, A., and Kroecker, T. S. (2017). Computational models of ethical decision-making: A coherence-driven reflective equilibrium model. *Cognitive Systems Research*, 46:61–74.

[541] Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12.

[542] Yoe, C. (2019). *Principles of risk analysis: decision making under uncertainty*. CRC press, Abingdon, UK.

[543] Yu, H., Shen, Z., C. Miao, C. L., Lesser, V. R., and Yang, Q. (2018). Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, pages 5527–5533, Stockholm, Sweden. International Joint Conferences on Artificial Intelligence.

[544] Yudkowsky, E. (2001). Creating friendly ai 1.0: The analysis and design of benevolent goal architectures. The Singularity Institute, San Francisco, USA.

[545] Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values*, 41(1):118–132.

[546] Zhang, R., McNeese, N. J., Freeman, G., and Musick, G. (2021). " an ideal human" expectations of ai teammates in human-ai teaming. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–25.

[547] Zhu, S. and Ma, W. (2015). Culture's influence on cockpit communication. In *Proceedings of the International Conference on Management, Computer and Education Informatization*, volume 25, pages 414–417. Citeseer.

[548] Złotowski, J., Proudfoot, D., Yogeeswaran, K., and Bartneck, C. (2015). Anthropomorphism: opportunities and challenges in human–robot interaction. *International journal of social robotics*, 7(3):347–360.

Curriculum Vitae

Suzanne Tolmeijer – CV

Date of Birth 2^{nd} July 1991

Personal Profile

Fourth year PhD student in Informatics with a passion for interdisciplinary projects. Background in socially-oriented artificial intelligence. Currently working on topics including machine ethics, trust in autonomous systems and diverse news recommender algorithms. Experience in teaching.

Education

September 2017 - February 2022
Doctoral program at the University of Zurich
Department of Informatics
Dynamic and Distributed Information Systems Group

Current topic: ethics and trust in autonomous systems.
Junior researcher representative within the faculty and university senate.
Responsible for news blog.
Teaching assistant.

September 2015 - June 2017
MSc Artificial lntelligence
VU University Amsterdam

Diploma cum Laude.
Thesis title: "An lntegrated Computational Phishing Detection Model"
Conducted at national research institute TNO
Teaching assistant

September 2012 - June 2015
BSc Lifestyle Informative
VU University Amsterdam

Diploma cum Laude with honours.
Thesis title: "How eHealth can improve Self Management in Cardiac Rehabilitation"
Conducted at ChipSoft
Teaching assistant

Teaching

2018 - 2021
Theses and Projects
University of Zurich
Under my guidance various students successfully completed
Five Master Tehses
One Master Project
Titles:

- "Application Prototype for Recognizing High Stress Situations"
- "Framework for News Recommendations"
- "Mobile Application for Aggregated News Recommendation"
- "Personalized News Recommendations"
- "When the Turing Test meets Trust"
- "Acceptance Autonomous Cars"