Year: 2022

Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making

Tolmeijer, Suzanne ; Christen, Markus ; Kandul, Serhiy ; Kneer, Markus ; Bernstein, Abraham

Abstract: While artificial intelligence (AI) is increasingly applied for decision- making processes, ethical decisions pose challenges for AI applica- tions. 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 collabora- tion? In this study, we investigate how the expert type (human vs. AI) and level of expert autonomy (adviser vs. decider) influence trust, perceived responsibility, and reliance. We find that partici- pants consider humans to be more morally trustworthy but less capable than their AI equivalent. This shows in participants' re- liance on AI: AI recommendations and decisions are accepted more often than the human expert's. However, AI team experts are per- ceived to be less responsible than humans, while programmers and sellers of AI systems are deemed partially responsible instead.

DOI: https://doi.org/10.1145/3491102.3517732

Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making

Suzanne Tolmeijer
University of Zurich
Switzerland

Markus Christen
University of Zurich
Switzerland

Serhiy Kandul
University of Zurich
Switzerland

Markus Kneer
University of Zurich
Switzerland

Abraham Bernstein
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 collaboration? In this study, we investigate how the expert type (human vs. AI) and level of expert autonomy (adviser vs. decider) 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.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Computer supported cooperative work**; **Empirical studies in collaborative and social computing**; • **Computing methodologies** → **Philosophical/theoretical foundations of artificial intelligence**.

KEYWORDS

Ethical AI, Trust, Responsibility, Human-AI Collaboration

ACM Reference Format:
Suzanne Tolmeijer, Markus Christen, Serhiy Kandul, Markus Kneer, and Abraham Bernstein. 2022. Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3491102.3517732

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 [111] and co-creation of art [80] to HR decisions [109] and clinical decision making [78, 142]. 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 [6]. Especially lay people seem to trust an AI more in various cases, such as forecasts of song popularity or romantic attraction [84]. 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 [28]. 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 [19]. All of the mentioned factors that can trigger algorithmic aversion depend on the decision domain and task type the AI performing in [21].

In this contribution, we compare user perception of AI vs. human involvement for tasks that require ethical decision making. While some tasks are generally accepted to be outsourced to AI completely, this is not the case for ethical decision making (e.g., [35, 117]). Rather, such tasks are usually expected to involve both humans and AI systems in a collaborative setting, where the AI could advice a human agent or the human could supervise the AI and intervene if necessary. The reason for this lies in the nature of ethical decision making, namely the question whether a ground truth in ethics exists and if so, what it should look like. Philosophers are divided over the question whether objective truth exists in ethics [47, 52, 88]. They are further divided over the question in virtue of what an action is to be assessed as right or wrong. Kant [69] famously placed strong emphasis on the agent's intentions, while consequentialists, such as Bentham [14] and Mill [96], tend to look more to outcomes. What is more, seemingly obvious desirable values can be somewhat inconsistent: maximizing equality can conflict with the maximization of individual liberty.

An example for this problem is how to implement different conceptions of fairness (e.g., procedural fairness and outcome fairness) into algorithmic decision making, as illustrated by the recent debate

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9157-3/22/04.
https://doi.org/10.1145/3491102.3517732

concerning the COMPAS recidivism algorithm [71]. It is mathematically impossible to adhere to all of people's different notions of fairness [74, 85]. Instead, the transparency that algorithms offer for discrimination and bias in decision making highlight the trade-offs between different values [73]. While research on implementing ethics in AI has being ongoing, it has been in a scattered and relatively limited fashion [133].

Part of what makes ethical AI so difficult to implement in practice, is the challenge of responsibility ascription — especially when a decision could lead to negative outcomes. The use of autonomous systems, for instance, could give rise to "responsibility gaps" — i.e. situations, where nobody can be held morally responsible [90]. In the context of ethical decision making for AI in severe contexts, such as with autonomous weapons systems, this has 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 [117]. The importance of the human element to ensure moral and legal accountability when using AI in security contexts is considered indispensable by stakeholders such as the ICRC [105]. In other words, there is a societal preference for letting a human be accountable for consequences of AI decisions at all times. Whether or not people perceive different parties involved in the AI system to be responsible is an ongoing topic of research. In addition to the theoretical discussion on moral accountability, there is the aspect of people's perceptions of moral responsibility in AI contexts. These perception are especially important for acceptance of autonomous AI practice [130]. Generally, users assign more responsibility to parties that have more autonomy in decision making [58]. Different types of agency lead to different responsibility ascriptions, such as to the AI artifact, the designer, and the user of the system [64]. The assigned responsibility also depends on the role and autonomy the AI has [81].

Assuming that humans need to be involved in ethical decision making, AI can be applied in a *human-in-the-loop* (HITL) setting or a *human-on-the-loop* (HOTL) setting [101]. 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. Given that human control over a system is not achieved by simply having human presence to authorise the use of force [95], 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 a (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 [77]. Chiang and Yin [24] 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 [3].

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 is gaining 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 collaboration, 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 and defense domain. Specifically, it focuses on trust placed in the AI and who is deemed responsible when humans and AI collaborate for ethical decision making. Given the current focus of AI for ethical decision making in the autonomous cars domain (e.g., [7, 18, 44, 83]), we focus on a different domain of unmanned aerial vehicles used in search and rescue as well as defence settings — domains where autonomous AI can be expected soon.

To this end, we 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) [25]. 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 be perceived to be clearly wrong. Since the Trolley Problem [128], 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 [13], 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 somewhat 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 collaboration and AI for ethical decision making, by presenting design implications of our findings.

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 [141]. 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 [119]. In this section, we summarize ongoing work, focusing on the current debate on ethical aspects of AI, the difference between trust and reliance, the difference in perception between humans and AI doing tasks, trust and perceived responsibility in AI.

2.1 Ethical Aspects of AI

AI is continuously being applied in more domains, including those that involve decisions with high ethical stakes such as in criminal law, health, or national security. In those contexts, decisions have an ethical component as they pertain to "behavior that is considered right, good, and proper" [65, p 3]. Given that people can disagree on what is considered 'right' behavior and whether an objective truth exists in ethics at all [47, 52, 88], many decisions in such contexts where AI could be involved have an ethical component. Take a medical AI system that support in diagnosis and treatment decisions as an example [30]: is it ethical to propose a treatment with a lower likelihood of succeeding but a higher quality of life and life expectancy? Given this pronounced potential of AI to impact ethical decision making in sensitive contexts, it is not surprising that many organizations have proposed guidelines for ethical AI. Examples include the "Recommendation of the Council on Artificial Intelligence" of the OECD [143], the "Recommendation on the human rights impacts of algorithmic systems" of the Council of Europe [126], and the "Ethics guidelines for trustworthy AI" of the European Commission [106]. Jobin et al. [63] have summarized the most common ethical principles present in those guidelines: transparency, justice and fairness, non-maleficence, responsibility, and privacy. One challenge related to these guidelines is their practical application. First, it can be hard to translate high-level concepts into something implementable for real-word cases [133], second, there can be trade-offs between these values [115] but the guidelines do not deal with the question how to address those, and third, even within each principle, it depends on the concrete definition whether a principle is met [93]. In summary, the guideline approach helps to raise awareness on ethical issues when using AI in sensitive domains, but its practical effect is limited.

Instead, for addressing ethical AI in real-world applications, two other approaches gain relevance, depending on the type of decisions made. One type of decision concerns situations, where AI makes decisions by its own due to practical constraints (e.g., because decision times are very short as for example in case of accidents in autonomous driving). For solving such problems, the field of 'machine ethics' is concerned with the question how to implement ethical theories in AI [5], such that ethically acceptable decisions result from AI systems operating autonomously. While there have been successful prototypes, this approach suffers from the problems sketched in the beginning of our contribution: there is no consensus in the discipline on which ethical theory should be used, how it should be interpreted, which algorithm works best, and how to properly evaluate the outcomes of AI deployed with ethical theory [133]. One suggestion for solving those problems is to determine the preferences of humans regarding the ethically best option in a decision problem, as for example the now famous moral machine has done so regarding autonomous cars behavior [7]. But it is highly disputed whether such a "majority ethics" would be in line with the protection of fundamental rights of different groups, including minorities.

A different type of decisions concern those where AI is not deciding autonomously, but where humans and AI systems collaborate for finding optimal solutions. This approach seems more fruitful for sensitive decision contexts where justification requirements have to be met, such as in triage decisions in medical emergencies [116]. Following the human-in-the-loop and human-on-the-loop distinction described above (which we indicate in this work as 'level of autonomy' of the expert), AI systems could either consult human deciders (e.g., to correct for known biases of humans in ethical decision making, [22]) or those systems are supervised by humans, such that they can intervene if the decision is considered to be unethical. For this approach, the rich literature on Value Sensitive Design (VSD) is relevant, a theoretically grounded approach to the design of technology that accounts for human values throughout the design process in a principled and comprehensive manner [41, 95]. For example, Cummings [27] combined ethics with VSD to achieve AI created using ethics-by-design. This approach of human-AI collaboration in ethical decision making has the benefit of being more practically applicable and increases the chances of technology uptake.

Related to the design of ethical AI for HCI practice, much emphasis has been placed on designing responsible [79] and fair [146] AI. Questions arise around which metaphors are appropriate to describe AI, what type of roles AI should take on, and whether they should mimic human-human interaction or not [138]. Especially in the context of designing AI as collaborators, there are many open questions, including which AI capabilities are needed and whether AI should be deceptive if it serves the greater good [137]. Design approaches should include awareness of the effect of cultural differences, privacy of users, consider accessibility, and potential environmental impact [72].

The end goal of all these approaches it to create AI that end users are willing to rely on. Reliance is directly related to a user's mental model on the capabilities of AI, since it influences a user's willingness to use and rely on the system [10]. One of the reasons users are not willing to rely on AI, is related to responsibility: the more responsible the user feels for an algorithm's outcome, the less likely they are to use it [112]. Another factor that influences reliance is trust. While some researchers claim reliance is the result of trust in the system, recent research has shown there can be a gap between reported trust and resulting reliance behavior [118] — an important reason for distinguishing between both variables in empirical work.

2.2 Trust versus Reliance

Trust is often researched in HCI, as it is said to increase acceptance and appropriate use of AI [97]. However, it is not always clear what is meant by trust, trustworthiness, reliance, and other related concepts [43]. An overview of Jacovi et al. [61] shows there are many types of trust that are not often specifically defined or distinguished. To highlight and better describe the difference between trust and reliance, we draw inspiration from how various fields deal with the two concepts.

Philosophic perspective. The philosophical literature focuses predominantly on questions regarding (i) the nature of trust, distrust, and trustworthiness, (ii) the conceptual relations between trust and related notions such as reliance, dependence, hope, and risk, (iii) the type of psychological attitude trust constitutes, (iv) conditions under which it is rational to trust, and the (v) the relation of trust and the will, i.e. whether one can (always) decide to trust [8, 9, 51, 53, 57, 91, 108, 110]. A core distinction in philosophy, which is often neglected in the empirical literature, regards trust and reliance. To rely on some entity or individual with respect to X is a matter of acting on the supposition that they will bring about X [57]. For instance, I might rely on my sprinkler system, or on my neighbor, to water the lawn. Trust, by contrast, is standardly limited to persons and, while it entails reliance, it goes beyond mere reliance. There are different ways to characterize the extra element which distinguishes trusting a person from relying on a person. On rationalist accounts (cf. e.g., [51, 110]), the fact that A trusts B with respect to X, generates reasons for B to bring about X. I might rely on you to water my hedge while on holiday, because otherwise, it quickly turns into an eyesore visible from your terrace. But if I trust you to do so, and make this manifest, you will have extra reasons to water the hedge related to my expectations, over and above those eyesore-related reasons you already have. In sum, while trust and reliance are related concepts, they should be considered separately.

Economics perspective. A field where reliance and trust are distinguished in empirical setting, is experimental economics. In this field, trust has been mainly conceptualized as the willingness to "put oneself in somebody else's hands" [4], for example to make one's payment in the game depend on the decision of others under conflict of interests, where the others (trustees) have an incentive to behave opportunistically, i.e. to take the action which would be detrimental for the trustor. Therefore, the utility of trustors depend on the likelihood of others to forgo the incentive to behave opportunistically, their trustworthiness. Reliance, on the other hand, does not require conflict of interest or the intentional resistance of the temptation to act opportunistically. This notion primarily refers to the principal's belief in the competence of the agent to perform the desired action, such as in the rate of unintentional errors or mistakes, which affect the expected outcome [45]. Uncertainty about the outcome is what trust and reliance have in common; the *source* of this uncertainty is where they differ: trust refers to the extent of opportunistic behavior by others and reliance refers to the extent of unintentional mistakes. In this sense, trust can bee seen as acceptance of strategic risk and reliance as acceptance of a lottery, a pre-defined risk of a failure [16, 17]. Interestingly, people's preferences regarding the source of uncertainty emerge even when the objective risk and expected outcome is the same [34, 45].

Responsibility bridges trust and reliance by reflecting the degree of one's influence on the outcome, conceptualized as the relative share of one's (intended) contribution to the probability of the outcome [11]. Responsibility attributions have been shown important in the broader context of ethical decision making. For example, when multiple agents are involved in the decision which harms others, individual wrongdoers are punished less [11]. The willingness to shift responsibility has been proposed as a motive to delegate ethically loaded decisions [50, 68]. Reduced responsibility is known to increase the rate of lying [139] and enhance antisocial behavior [1].

HCI perspective. In the context of HCI, an often used definition of trust is "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*" [77, p 51]. Trust in the system influences whether and how AI is used in practice, which in turn is highly relevant when designing AI. For example, how the AI and its capabilities are framed by the designer highly impact acceptance and accuracy perceptions of users [76]. While trust has been established as an influential factor in technology use, it has only been fairly recently that methods have emerged to systematically include trust insights in system design [120]. The research on explainable AI, which attempts to find user-friendly ways of opening up the 'black box' of deep learning systems, is an example of how HCI researchers are attempting to achieve appropriate user trust by influencing mental models [2].

Reliance on the other hand has mostly been investigated in the context of over- and under-reliance. Systems should be designed to instigate appropriate reliance — that is, users should not overtrust the system to expect more than the system is capable of, nor distrust it when they expect less than what is possible [77]. Incorrectly calibrated trust in AI, which can for example be caused by encountering mistakes early in the usage process, unjustly lowers user reliance [104, 132]. The large effect of system framing and design on user trust and reliance also becomes relevant in the context of ethical AI: through design, HCI researcher can help the user calibrate their expectations of the system, leading to appropriate trust and reliance.

In conclusion, both trust and reliance are relevant in the context of user experience of AI. However, while they are related, they should be treated and measured as independent concepts. Implications for HCI research include that different trust related concepts and its prerequisites should be clearly defined and distinguished. In this paper, we therefore define trust as the belief that "an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability" [77, p 51], while reliance is defined as "a discrete process of engaging or disengaging"[77, p 50] with the AI system.

2.3 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 [142], group decision making [70], and advise on what to eat [127] or watch [99]. Now, applications are moving towards autonomous analysis of tasks, such as diagnosis based on medical images [49] or autonomous task execution like driving a car [60]. Within 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 [46].

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 an 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 [35]. 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 [6]. On an individual user level, Chen et al. [23] 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 [102]. Human artwork is evaluated more highly than AI artwork [113]. On the other hand, news articles written by AI and human news editors were considered equally credible [140].

One important factor that relates to perception of AI is trust. When trust in AI systems is higher than in human experts, this can lead to what Logg et al. [84] 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. [129] 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 [6]. One possible explanation is the machine heuristic, in which humans consider AI to be more objective and less ideology-biased than humans [122]. 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 without a justified reason, 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 [28]. One way to overcome this aversion, is by framing the system to be a learning system [15]. In a literature review, Jussupow et al. [66] 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 [21]. 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.

2.4 Perceived Responsibility of Autonomous AI

Part of the challenge of using autonomous AI is the ascription of responsibility of decision making. Yet, responsibility is fundamental for autonomy of AI systems [29]. 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. [33] 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 [75]. 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 [82]. In some cases, such as by younger adults, blame falls more on the programmer making the AI rather than the AI itself [40]. 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*" [54, p 14]. This effect of responsibility diffusion has been researched in social psychology (e.g., [38, 100, 136]), but not yet in the contest of human-AI collaboration.

In addition to issues with perception of responsibility, the legal system is not equipped to deal with criminal liability of AI systems yet [107]. For instance, liability and data usage of AI creating news articles are currently becoming an issue [98]. 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.

In addition, we expect the level of autonomy to have an influence on responsibility perception, as a higher level of autonomy implies higher decision power, which is often linked to responsibility [31]. 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 [114, 121]. 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 collaboration, 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 [48, 82, 135]. 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.

2.5 Human-AI Collaboration 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 collaboration could be a fruitful alternative to combine strengths and produce new possibilities for the future of work [62]. 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 collaborative effort that includes humans as well. In the context of the role of humans in human-AI collaboration, two prominent configurations of human-AI collaboration have been discussed: human-in-the loop and human-on-the loop settings (see, e.g, [36, 59, 101]). 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).

The varying degree of human involvement in human-AI collaboration might impact people's perception of AI, and therefore affect the degree of people's reliance on AI. 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.

3 METHOD

To study our three research questions and hypotheses, 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). The domain of unmanned aerial vehicles was chosen to expand the current focus on autonomous cars (e.g., [7, 18, 44, 83]) and broaden the discussion of possible applications of AI for ethical decision making.

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[1] 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).

3.1 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 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 [13]. These scenarios were developed together with domain experts from the Swiss Center for Drones and Robotics of the Department of Defense, Civil Protection and Sport (DDPS) to ensure relevance and sufficient realism.

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)[2].

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 images 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.

[1] prolific.co

[2] https://osf.io/v9w8a/?view_only=400fd838cde6447b83af210b7a237295

	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.**

Table 1: Scenario types

3.3 Main Experiment

3.3.1 Participants. Participants were recruited from the general population, since 1) this allows for comparison of ethical preferences against studies in other domains, and 2) support of the general population will be needed before governments can consider outsourcing ethical decisions to AI. 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 were female (253), 39% were 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.

3.3.2 Design. We used a 2x2 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 2.

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) [89]. 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 [89] 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 [92], affinity for technology interaction [39] and utilitarian scale [67]). Furthermore, also the differentiation between "maximizing lives saved" and "minimizing lives lost" served as control condition.

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. [103] 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 1).

The procedure for the experiment can be found in Figure 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 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) [32, 56, 134]. In order to take these potential differences into account, the experts provided two times an advice that maximized probability and 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 limit expert performance as a possible confounding variable. We will discuss the impact of this design choice in our discussion section.

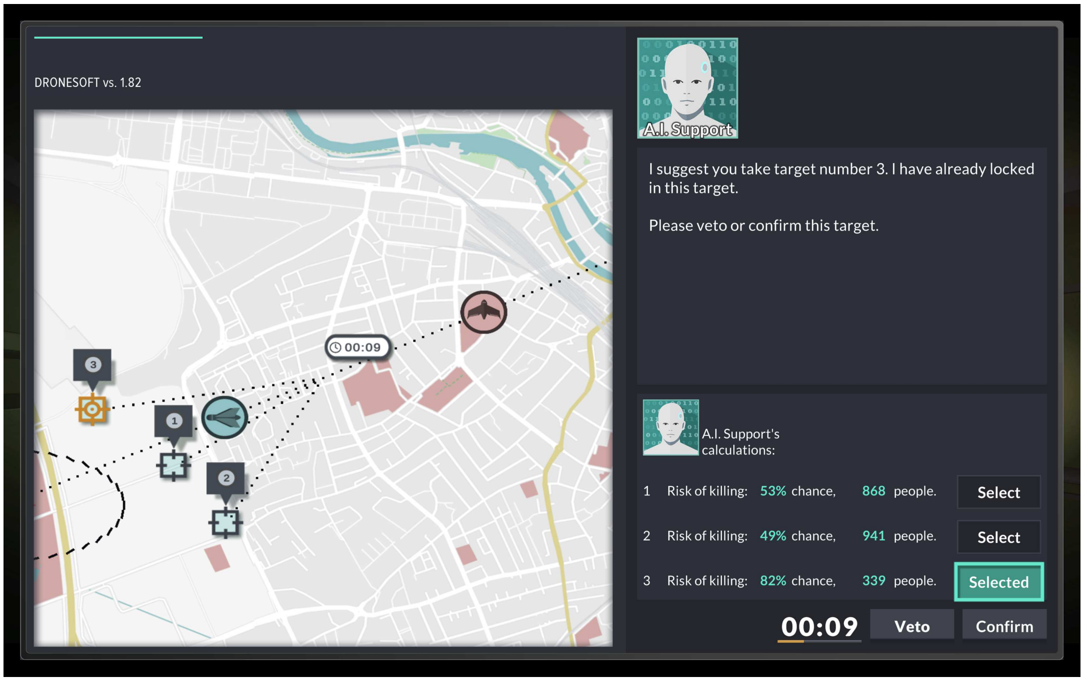
Finally, after the missions, we controlled how serious the participants took the scenarios with two engagement questions. In this post-test phase, we also measured their affinity with technology interaction and utilitarian preference as a control variable. The trust scale was presented for the AI and 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 as well as the full dialog flow of the system can be found in the provided OSF link.

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.

veto kill.png



I suggest you take target number 3. I have already locked in this target.

Please veto or confirm this target.

A.I. Support's calculations:

1 Risk of killing: 53% chance, 868 people.

2 Risk of killing: 49% chance, 941 people.

3 Risk of killing: 82% chance, 339 people.

Figure 1: Screenshot of a decision during the simulation for a human-on-the-loop setting and defense scenario. For the decisions, the left part of the screen showed the possible crash sights, while the upper right corner showed the expert's opinion. The blue icons is the drone being operated, the red icon is an income bomb.

Participants had to select their choice in the bottom right.

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. 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 expert autonomy yielded an effect of F(1,461) = 2.2, p = 0.136, η^2=0.005, and an effect of F(1,461) = 0.3, p = 0.533, η^2<0.001 for AI and human experts respectively. 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(1,461) = 0.4, p = 0.551, η^2<0.001 and F(1,461) = 1.0, p = 0.330, η^2=0.002 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(1,461) = 1.8, p = 0.186, η^2=0.004 and F(1,461) = 0.2, p = 0.613, η^2<0.001 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, d = 0.24. 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, d = 0.52. 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, d = 0.30. While moral trust and overall trust show a smaller effect size, capacity trust displays a medium effect size.

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 number of times this happened for each trust item in the human and AI expert setting. There was a significant difference in number of items labeled not applicable between the human expert (M = 51.2, SD = 18.5) and the AI (M = 71.6, SD = 28.3); t(14) = 5.06, p = p < 0.001, d = 1.2. 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 (M = 20.5, SD = 13.8) and AI experts (M = 10.6, SD = 7.1) results in a significant effect: t(7) = 2.66, p = 0.0326, d = 0.84. Moral trust is also assigned significantly less to AI (M = 120.8, SD = 12.6) then to the human expert (M = 44.5, SD = 9.3): t(7) = 16.5, p < 0.001, d = 6.55.

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

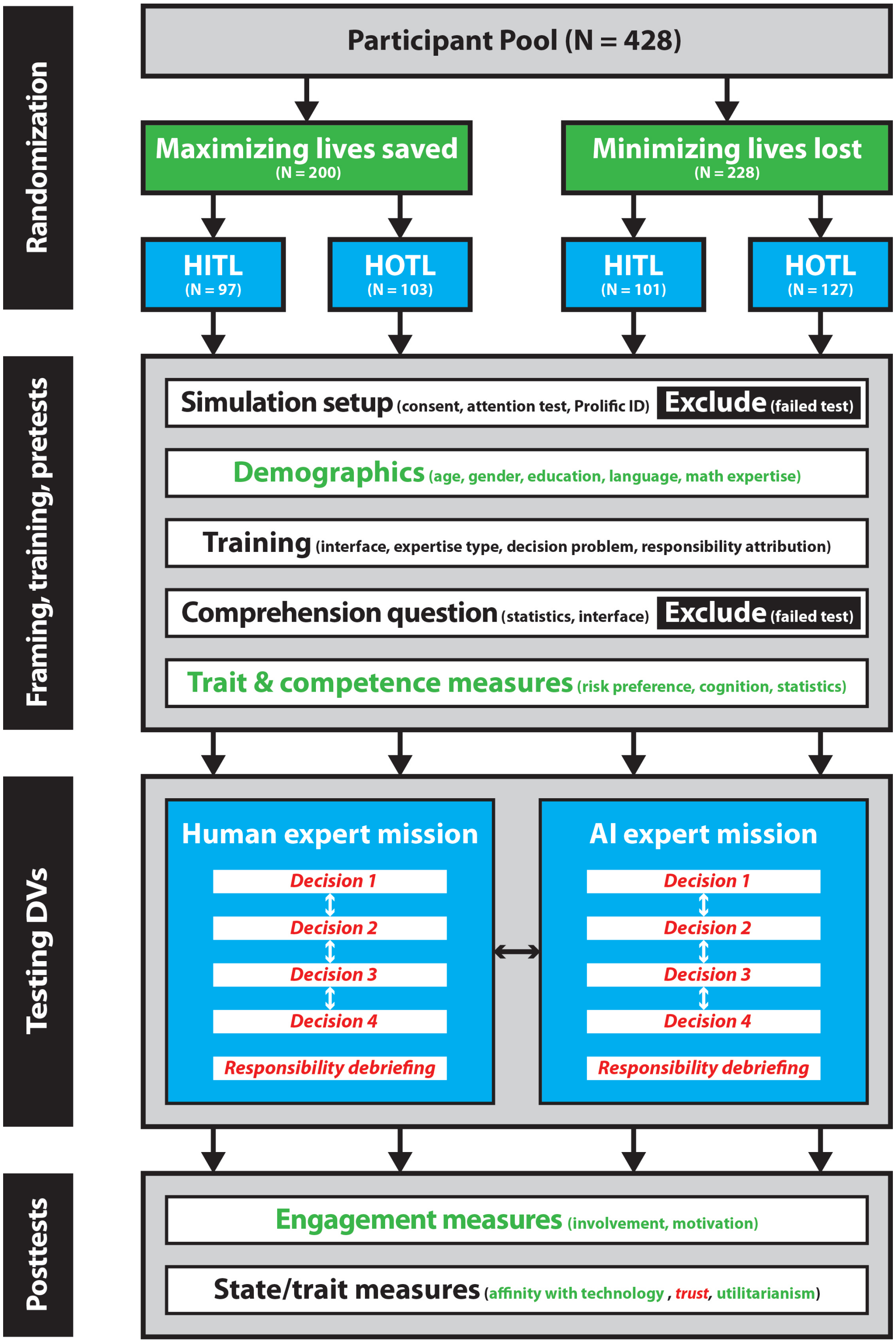

Randomization
Framing, training, pretests
Testing DVs
Posttests
Participant Pool (N = 428)
Maximizing lives saved (N = 200)
Minimizing lives lost (N = 228)
HITL (N = 97)
HOTL (N = 103)
HITL (N = 101)
HOTL (N = 127)
Simulation setup (consent, attention test, Prolific ID)
Exclude (failed test)
Demographics (age, gender, education, language, math expertise)
Training (interface, expertise type, decision problem, responsibility attribution)
Comprehension question (statistics, interface)
Exclude (failed test)
Trait & competence measures (risk preference, cognition, statistics)
Human expert mission
Decision 1
Decision 2
Decision 3
Decision 4
Responsibility debriefing
AI expert mission
Decision 1
Decision 2
Decision 3
Decision 4
Responsibility debriefing
Engagement measures (involvement, motivation)
State/trait measures (affinity with technology, trust, utilitarianism)

Figure 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.

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 slightly more than the human expert. They have a higher capacity trust in AI, while having a somewhat 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, implying that both algorithmic appreciation and algorithmic aversion are displayed but for different dimensions.

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, η^2=0.001 and F(1,461) = 1.63, p = 0.203, η^2=0.004 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, η^2=0.01, indicating a small 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 small but significant interaction for the programmer's responsibility between expert autonomy and framing of the ethical scenario (F(1.461) = 6.55, p = 0.0108, η^2=0.01), as well as between expert autonomy and mission order (F (1,461) = 4.37, p = 0.0372, η^2=0.01). 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 deemed 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 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(936) = 0.08, p = 0.940, d = 0.005). 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, d = 1.0.). 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, d = 0.38. 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, d = 0.06). 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.

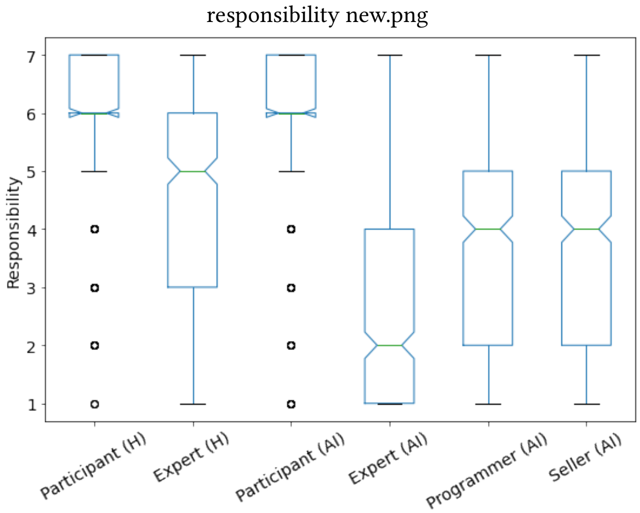


Figure 3: Boxplot of the assigned responsibility scores. The 'notch' around the median shows a 95% confidence interval of the median.

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. In the context of the given tasks, 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 influences responsibility perceptions for the programmer, and had an interaction with the framing of the scenarios and order of presented experts. However, level of autonomy did not have an influence on perception for other responsibility perceptions. This partially confirms H2: AI is perceived to be less responsible that a human expert.

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 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.).

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

Table 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.

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 slightly 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 as often on AI as on the human expert and showed no algorithmic aversion.

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.

5.1 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 slightly higher moral trust scores assigned to humans experts, but also in the significant amount of times participants felt items of moral trust did not fit the AI. The fact that participants seems to either think AI is less morally trustworthy, or AI is not even *able* to be morally trustworthy, has strong implications for AI making ethical decisions. A plausible interpretation of the result is that the human decider in our experiments do not consider the AI system to be a moral agent, in contrary to the human expert, although both entities acted in an interchangeable way (i.e., they provided the same recommendation or made the same 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 somewhat 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. Despite the fact that the found results had different effect sizes, the stability of these findings seem to point to a set expectation of what humans and AI can be trusted to do, independently of how they are deployed.

A possible explanation lies in the realm of heuristics and cognitive bias, and their influence on trust formation in autonomous technology, which is an ongoing topic of research (e.g., [12, 26, 87, 123, 124]) and relates to algorithmic aversion and appreciation. For example, when a situation is less predictable [125] or a situation has more uncertainty [26], users tend to overtrust AI. Starting expectations, which can be shaped by many things including earlier technology interactions [55] and even science fiction [35], influence the trust that develops — better performance than the user expected can lead to increased trust and vice verse [145]. The design of systems can lead to overtrust [87] and in some cases, people display the machine heuristic where they trust a machine more than a human [123]. These heuristics, which often manifest stronger under time pressure [131] like in our experiment, can also lead to a apparent paradox in perception. It has already been found that implicit attitudes towards AI can differ from explicit propensity to trust [94]. Recently, this gap has specifically been found for AI decision making tools [118].

Algorithmic aversion is often linked with system errors [28] and our AI did not make explicit mistakes. Additionally, lay people often trust AI more than experts would [84]. This combination can explain the higher capacity trust participants experienced. However, this same effect was not present for moral trust. This could possibly be explained through the philosophical perspective that somewhat connects trust to human agency. In that spirit, we could interpret that whilst AIs can have a capacity for trustworthiness, their lack of human agency denies them a moral trustworthiness. Hence, the distinction would become one of ontological framing by the users.

For the CHI community, in general, it is pertinent to get a better understanding of user's expectations and tendencies, to be able to design a system that does not invoke unwanted bias. In the context of ethical decision making, where outcomes can be quite severe, this becomes even more important.

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 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.

5.3 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 marginally more often on AI advice 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 not less 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 comparable reliance. The algorithmic appreciation caused by a lack of mistakes seems to weight stronger than the lack of ascribed moral agency. One explanation for the higher capacity trust and reliance can be the earlier mentioned 'machine heuristic' [122]. Possibly, participants consider the AI to be somewhat more objective and less ideology-driven, also in an ethical decision making setting.

Like trust, reliance in our setting is also likely to be influenced by heuristics, especially since there is less information available [86], and limited time to process information [131]. Other known triggers of heuristics leading to reliance differences range from errors in the system [28] to design and looks of the system [42] and past experience [132] — the first interactions set positive expectations for the remaining interactions.

Moreover, in our setting, participants could follow two basic strategies: expected utility maximization or risk optimization, which correspond to the two types of advice produced by the expert. Future research should investigate the link between time pressure, system features and advice types/strategies on reliance.

5.4 Expert autonomy

In our experiment, the level of autonomy, HITL vs. HOTL, did not significantly affect the trust and reliance. This can be explained by the fact that in our scenarios the change in autonomy is subtle: it is always the participant who is (reminded to be fully) responsible for the decision. Both types of framing set a default option and asked participants to actively choose another option in case they disagree with the system ("opt-out" design). This design was chosen in deliberation with domain experts, since the domain is not likely to get fully autonomous AI for this type of application. However, we believe that larger differences in autonomy of the expert, for example, going from human-in-the-loop to human-out-of-the loop design, could substantially affect participants' perceptions of human and AI experts. Given that autonomy can be more complex that 'just' HITL and HOTL [95], future work is needed on the different possible autonomy levels.

5.5 Design implications for ethical AI

In sum, we find that participants had slightly 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 collaboration. 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 users show for a human vs. AI expert.

One conclusion of this finding that the control conditions of AI may be of less relevance than expected is that the focus on human-AI collaboration should be less on control and more on how the involvement of AI improves human ethical decision making. An important factor in that respect will be the time available for actual decision making: if time is short, AI advice or decisions should make clear which value was guiding in the decision process (e.g., maximizing the expected number of people to be saved irrespective of any characteristics of the individuals involved), such that the human decider can make (or evaluate) the decision in an ethically informed way. If time for deliberation is available, a AI decision support system could be designed in a way to counteract human biases in ethical decision making (e.g., point to the possibility that human deciders solely focus on utility maximization and in this way neglecting fundamental rights of individuals) such that those biases can become part of the deliberation process.

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 and make sure only appropriate decision heuristics are triggered.

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 Limitations of the study

Several study limitations are worth mentioning. First, we rely on non-expert subjects. In general, lay people are more likely to show algorithmic appreciation than experts [84], since they know less of the domain. However, since our scenario lacks ground truth [144], it is difficult to predict how the choices of domain experts, in particular with respect to moral responsibility, would differ from our non-expert subjects. It is, however, worth noticing that understanding non-experts' preferences in these scenarios might be a worthy goal in itself, as 1) such drone applications can only be implemented if there is not a large public backlash, and 2) they are increasingly likely to encounter moral decision problems with AI support in the future without much training. We plan to extend this work to compare lay people's perceptions against domain experts.

Second, our scenarios rely on a modified trolley dilemma. However, a deviation from the original dilemma is that the baseline, the damages from failing to take an action, is vaguely specified. This was intentional, to prevent an unwanted learning effect during the missions. In a more realistic setting, users will likely have more explicit information about the consequences of their actions and inaction.

Third, the scenario's drone-based context in the military or search and rescue domain might be influencing some subjects. However, most moral decision dilemmas are likely to evoke an emotional response [20]. Since our drone scenario include generic features of ethical dilemmas (such as the distinction between causing or preventing harm [37]) with the main goal to reduce or prevent casualties, we believe that our results can be compared with those of other relevant studies on ethical decisions with AI systems. However, more research is needed in different domains than autonomous vehicles and drones, to investigate to what extent user preferences generalize across domains.

Fourth, 50% of the advice provided per mission minimized risk, 50% maximized expected utility. The order of advice type was randomized. Further analyses could be conducted to explore whether participants' individual preferences as to what constitutes the morally best outcome had an impact on advice reliance, and whether the latter interacted with the order of advice type within missions.

Finally, not only advice type, but also the difficulty of clearly distinguishing between preferred and non-preferred options could influence reliance behavior. Given that the task of calculating, e.g., expected utilities is non-trivial (as participants need to not only account for probabilities and casualties of a single option, but of all options), some missions could be harder to evaluate compared to others.

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 somewhat 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.

ACKNOWLEDGMENTS

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). This work is partially funded by National Research Program (NRP) 77 'Digital Transformation' of the Swiss National Science Foundation under project number 407740_187494. We thank Koboldgames for their support during the design process of the simulation.

REFERENCES

[1] Klauw Abbink and Benedikt Herrmann. 2011. The Moral Costs of Nastiness. *Economic Inquiry* 49, 2 (2011), 631–633.

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–18.

[3] Nadia Adnan, Shahrina Md Nordin, Mohamad Ariff bin Bahruddin, and Murad Ali. 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 (2018), 819–836.

[4] Carlos Alós-Ferrer and Federica Farolfi. 2019. Trust Games and Beyond. *Frontiers in Neuroscience* 13 (2019), 887.

[5] Michael Anderson, Susan Leigh Anderson, and Chris Armen. 2004. Towards machine ethics. In *AAAI-04 workshop on agent publishers: theory and practice, San Jose, CA*. AAAI Press, Phoenix, Arizona, USA, 2–7.

[6] Theo Araujo, Natali Helberger, Sanne Kruikemeier, and Claes H De Vreese. 2020. In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35, 3 (2020), 611–623.

[7] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The moral machine experiment. *Nature* 563, 7729 (2018), 59–64.

[8] Annette Baier. 1986. Trust and antitrust. *ethics* 96, 2 (1986), 231–260.

[9] Judith Baker. 1987. Trust and rationality. *Pacific philosophical quarterly* 68, 1 (1987), 1–13.

[10] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. The AAAI Press, Palo Alto, California USA, 2–11.

[11] BJÖRN BARTLING and URS FISCHBACHER. 2012. Shifting the Blame: On Delegation and Responsibility. *The Review of Economic Studies* 79, 1 (2012), 67–87.

[12] Ghassan F Bati and Vivek K Singh. 2018. "Trust Us" Mobile Phone Use Patterns Can Predict Individual Trust Propensity. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14.

[13] Christopher W Bauman, A Peter McGraw, Daniel M Bartels, and Caleb Warren. 2014. Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass* 8, 9 (2014), 536–554.

[14] Jeremy Bentham. 1996. *The collected works of Jeremy Bentham: An introduction to the principles of morals and legislation.* Clarendon Press, Oxford, United Kingdom.

[15] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. 2021. Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn. *Business & Information Systems Engineering* 63, 1 (2021), 55–68.

[16] Iris Bohnet, Fiona Greig, Benedikt Herrmann, and Richard Zeckhauser. 2008. Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review* 98, 1 (March 2008), 294–310.

[17] Iris Bohnet and Richard Zeckhauser. 2004. Trust, risk and betrayal. *Journal of Economic Behavior & Organization* 55, 4 (2004), 467–484.

[18] Jean-François Bonnefon, Azim Shariff, and Iyad Rahwan. 2019. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proc. IEEE* 107, 3 (2019), 502–504.

[19] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239.

[20] Sidney Callahan. 1988. The role of emotion in ethical decisionmaking. *Hastings Center Report* 18, 3 (1988), 9–14.

[21] Noah Castelo, Maarten W Bos, and Donald R Lehmann. 2019. Task-dependent algorithm aversion. *Journal of Marketing Research* 56, 5 (2019), 809–825.

[22] Danton S Char, Nigam H Shah, and David Magnus. 2018. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine* 378, 11 (2018), 981.

[23] Jin Chen, Cheng Chen, Joseph B. Walther, and S. Shyam Sundar. 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*. Association for Computing Machinery, New York, NY, USA, Article 299, 7 pages. https://doi.org/10.1145/3411763.3451735

[24] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *13th ACM Web Science Conference 2021*. Association for Computing Machinery, New York, NY, USA, 120–129.

[25] Markus Christen, Darcia Narvaez, Julaine D Zenk, Michael Villano, Charles R Crowell, and Daniel R Moore. 2021. Trolley dilemma in the sky: Context matters when civilians and cadets make remotely piloted aircraft decisions. *PLoS one* 16, 3 (2021), e0247273.

[26] Dan Conway, Fang Chen, Kun Yu, Jianlong Zhou, and Richard Morris. 2016. Misplaced Trust: A Bias in Human-Machine Trust Attribution–In Contradiction to Learning Theory. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 3035–3041.

[27] Mary L Cummings. 2006. Integrating ethics in design through the value-sensitive design approach. *Science and engineering ethics* 12, 4 (2006), 701–715.

[28] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[29] Virginia Dignum. 2018. Ethics in artificial intelligence: introduction to the special issue. , 3 pages.

[30] Steven E Dilsizian and Eliot L Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports* 16, 1 (2014), 441.

[31] Raymond Duch, Wojtek Przepiorka, and Randolph Stevenson. 2015. Responsibility attribution for collective decision makers. *American Journal of Political Science* 59, 2 (2015), 372–389.

[32] Gerd Gigerenzer Eduard Brandstaetter and Ralph Hertwig. 2006. The Priority Heuristic: Making Choices Without Trade-Offs. *Psychological Review* 113, 2 (2006), 409–462.

[33] Ziv Epstein, Sydney Levine, David G Rand, and Iyad Rahwan. 2020. Who gets credit for AI-generated art? *Iscience* 23, 9 (2020), 101515.

[34] Mike Farjam. 2019. On whom would I want to depend; humans or computers? *Journal of Economic Psychology* 72 (2019), 219–228.

[35] Ethan Fast and Eric Horvitz. 2017. Long-term trends in the public perception of artificial intelligence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 31. The AAAI Press, Palo Alto, California USA, 963–969.

[36] Joel E. Fischer, Chris Greenhalgh, Wenchao Jiang, Sarvapali D. Ramchurn, Feng Wu, and Tom Rodden. 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 (2021), e4082. https://doi.org/10.1002/cpe.4082 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.4082 e4082 cpe.4082.

[37] Philippa Foot. 2002. *Moral Dilemmas: and other topics in moral philosophy.* Clarendon Press, Oxford, United Kingdom.

[38] Donelson R Forsyth, Linda E Zyzniewski, and Cheryl A Giammanco. 2002. Responsibility diffusion in cooperative collectives. *Personality and Social Psychology Bulletin* 28, 1 (2002), 54–65.

[39] Thomas Franke, Christiane Attig, and Daniel Wessel. 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 (2019), 456–467.

[40] Nathan G. Freier, Elia J. Nelson, Amanda Rotondo, and Wai Kay Kong. 2009. *The Moral Accountability of a Personified Agent: Young Adults' Conceptions.* Association for Computing Machinery, New York, NY, USA, 4609–4614. https://doi.org/10.1145/1520340.1520708

[41] Batya Friedman, Peter Kahn, and Alan Borning. 2002. *Value sensitive design: Theory and methods.* Technical Report 2-12. University of Washington.

[42] Anna-Katharina Frison, Philipp Wintersberger, Andreas Riener, Clemens Schartmüller, Linda Ng Boyle, Erika Miller, and Klemens Weigl. 2019. In UX We Trust: Investigation of Aesthetics and Usability of Driver-Vehicle Interfaces and Their Impact on the Perception of Automated Driving. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.

[43] Felix Gille, Anna Jobin, and Marcello Ienca. 2020. What we talk about when we talk about trust: theory of trust for AI in healthcare. *Intelligence-Based Medicine* 1 (2020), 100001.

[44] Jan Gogoll and Julian F Müller. 2017. Autonomous cars: in favor of a mandatory ethics setting. *Science and engineering ethics* 23, 3 (2017), 681–700.

[45] Jan Gogoll and Matthias Uhl. 2018. Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics* 74 (2018), 97–103.

[46] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. 2018. When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research* 62 (2018), 729–754.

[47] Peter A Graham. 2010. In defense of objectivism about moral obligation. *Ethics* 121, 1 (2010), 88–115.

[48] Nina Grgić-Hlača, Christoph Engel, and Krishna P. Gummadi. 2019. Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 178 (Nov. 2019), 25 pages. https://doi.org/10.1145/3359280

[49] Holger A Haenssle, Christine Fink, Roland Schneiderbauer, Ferdinand Toberer, Timo Buhl, Andreas Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, 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 (2018), 1836–1842.

[50] John R. Hamman, George Loewenstein, and Roberto A. Weber. 2010. Self-Interest through Delegation: An Additional Rationale for the Principal-Agent Relationship. *The American Economic Review* 100, 4 (2010), 1826–1846.

[51] Russell Hardin. 1993. The street-level epistemology of trust. *Politics & society* 21, 4 (1993), 505–529.

[52] Gilbert Harman. 1975. Moral relativism defended. *The Philosophical Review* 84, 1 (1975), 3–22.

[53] Katherine Hawley. 2014. Trust, distrust and commitment. *Noûs* 48, 1 (2014), 1–20.

[54] Maria Hedlund and Erik Persson. 2021. Expert responsibility in AI development. *Proceedings of the International Conference of Public Policy (ICPP5)* 5 (2021), 1–24.

[55] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[56] Charles A. Holt and Susan K. Laury. 2002. Risk Aversion and Incentive Effects. *American Economic Review* 92, 5 (December 2002), 1644–1655.

[57] Richard Holton. 1994. Deciding to trust, coming to believe. *Australasian journal of philosophy* 72, 1 (1994), 63–76.

[58] Joo-Wha Hong and Dmitri Williams. 2019. Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. *Computers in Human Behavior* 100 (2019), 79–84.

[59] Michael Horowitz and Paul Scharre. 2015. *An introduction to autonomy in weapon systems.* Technical Report. Center for A New American Security.

[60] Rasheed Hussain and Sherali Zeadally. 2018. Autonomous cars: Research results, issues, and future challenges. *IEEE Communications Surveys & Tutorials* 21, 2 (2018), 1275–1313.

[61] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. Association for Computing Machinery, New York, NY, USA, 624–635.

[62] Mohammad Hossein Jarrahi. 2018. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons* 61, 4 (2018), 577–586.

[63] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[64] Deborah G Johnson and Mario Verdicchio. 2019. AI, agency and responsibility: the VW fraud case and beyond. *Ai & Society* 34, 3 (2019), 639–647.

[65] Michael S Josephson and Wes Hanson. 2002. *Making ethical decisions.* Josephson Institute of ethics Marina del Rey, CA, Los Angeles, CA, USA.

[66] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. 2020. Why are we Averse towards Algorithms? A Comprehensive Literature Review on Algorithmic Aversion. In *Proceedings of the 28th European Conference on Information Systems (ECIS).* ECIS, Marrakech, Morocco, 1–16.

[67] Guy Kahane, Jim AC Everett, Brian D Earp, Lucius Caviola, Nadira S Faber, Molly J Crockett, and Julian Savulescu. 2018. Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review* 125, 2 (2018), 131.

[68] Serhiy Kandul and Oliver Kirchkamp. 2018. Do I care if others lie? Current and future effects when lies can be delegated. *Journal of Behavioral and Experimental Economics (formerly The Journal of Socio-Economics)* 74, C (2018), 70–78.

[69] Immanuel Kant. 1785. *Groundwork of the metaphysics of morals.* Cambridge, Cambridge, United Kingdom.

[70] Nikos I Karacapilidis and Costas P Pappis. 1997. A framework for group decision support systems: Combining AI tools and OR techniques. *European Journal of Operational Research* 103, 2 (1997), 373–388.

[71] Aria Khademi and Vasant Honavar. 2020. Algorithmic bias in recidivism prediction: A causal perspective (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. The AAAI Press, Palo Alto, California USA, 13839–13840.

[72] Lorraine Kisselburgh, Michel Beaudouin-Lafon, Lorrie Cranor, Jonathan Lazar, and Vicki L Hanson. 2020. HCI Ethics, Privacy, Accessibility, and the Environment: A Town Hall Forum on Global Policy Issues. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–6.

[73] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (2018), 113–174.

[74] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017) (Leibniz International Proceedings in Informatics (LIPIcs), Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 43:1–43:23.

[75] Markus Kneer and Michael T Stuart. 2021. Playing the Blame Game with Robots. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction.* Association for Computing Machinery, New York, NY, United States, 407–411.

[76] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. *Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems.* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[77] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[78] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. 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* (Yokohama, Japan) *(CHI '21).* Association for Computing Machinery, New York, NY, USA, Article 392, 14 pages. https://doi.org/10.1145/3411764.3445472

[79] Min Kyung Lee, Nina Grgić-Hlača, Michael Carl Tschantz, Reuben Binns, Adrian Weller, Michelle Carney, and Kori Inkpen. 2020. Human-centered approaches to fair and responsible AI. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–8.

[80] Zhuying Li, Yan Wang, Wei Wang, Stefan Greuter, and Florian 'Floyd' Mueller. 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* (Honolulu, HI, USA) *(CHI EA '20).* Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3382976

[81] Gabriel Lima and Meeyoung Cha. 2020. Descriptive AI Ethics: Collecting and Understanding the Public Opinion, In Ethics in Design Workshop. *arXiv preprint arXiv:2101.05957* 1, 1–6.

[82] Gabriel Lima, Nina Grgić-Hlača, and Meeyoung Cha. 2021. 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.* Association for Computing Machinery, New York, NY, USA, 1–17.

[83] Patrick Lin. 2016. *Why ethics matters for autonomous cars.* Springer, Berlin, Heidelberg, Heidelberg, Germany, 69–85.

[84] Jennifer M Logg, Julia A Minson, and Don A Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (2019), 90–103.

[85] Michele Loi and Markus Christen. 2019. How to include ethics in machine learning research. *ERCIM News* 116, 3 (2019), 5.

[86] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–16.

[87] Stefanie M. Faas, Johannes Kraus, Alexander Schoenhals, and Martin Baumann. 2021. Calibrating Pedestrians' Trust in Automated Vehicles: Does an Intent Display in an External HMI Support Trust Calibration and Safe Crossing Behavior?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.* Association for Computing Machinery, New York, NY, USA, 1–17.

[88] John Mackie. 1990. *Ethics: Inventing right and wrong.* Penguin UK, London, United Kingdom.

[89] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction.* Elsevier, Amsterdam, the Netherlands, 3–25.

[90] Andreas Matthias. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and information technology* 6, 3 (2004), 175–183.

[91] Victoria McGeer. 2008. Trust, hope and empowerment. *Australasian Journal of Philosophy* 86, 2 (2008), 237–254.

[92] Ree M Meertens and Rene Lion. 2008. Measuring an individual's tendency to take risks: the risk propensity scale 1. *Journal of Applied Social Psychology* 38, 6 (2008), 1506–1520.

[93] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[94] Stephanie M Merritt, Heather Heimbaugh, Jennifer LaChapell, and Deborah Lee. 2013. I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated system. *Human factors* 55, 3 (2013), 520–534.

[95] Leila Methnani, Andrea Aler Tubella, Virginia Dignum, and Andreas Theodorou. 2021. Let Me Take Over: Variable Autonomy for Meaningful Human Control. *Frontiers in Artificial Intelligence* 4 (2021), 133.

[96] John Stuart Mill. 1861. *1998. Utilitarianism, edited with an introduction by Roger Crisp.* Oxford University Press, New York, NY, USA.

[97] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[98] Matteo Monti. 2019. Automated journalism and freedom of information: Ethical and juridical problems related to AI in the press field. *Opinio Juris in Comparatione* 1 (2019), 2018.

[99] Rajatish Mukherjee, Gerdur Jonsdottir, Sandip Sen, and Partha Sarathi. 2001. Movies2go: an online voting based movie recommender system. In *Proceedings of the fifth international conference on Autonomous agents*, Vol. 5. Association for Computing Machinery, New York, NY, United States, 114–115.

[100] Clifford Mynatt and Steven J Sherman. 1975. Responsibility attribution in groups and individuals: A direct test of the diffusion of responsibility hypothesis. *Journal of Personality and Social Psychology* 32, 6 (1975), 1111.

[101] Saeid Nahavandi. 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 (2017), 10–17.

[102] David T. Newman, N. Fast, and Derek Harmon. 2020. When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160 (2020), 149–167.

[103] Evangelos Niforatos, Adam Palma, Roman Gluszny, Athanasios Vourvopoulos, and Fotis Liarokapis. 2020. *Would You Do It?: Enacting Moral Dilemmas in Virtual Reality for Understanding Ethical Decision-Making.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3313831.3376788

[104] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *26th International Conference on Intelligent User Interfaces.* Association for Computing Machinery, New York, NY, USA, 340–350.

[105] International Review of the Red Cross. 2019. *Artificial intelligence and machine learning in armed conflict: A human-centred approach.* Technical Report 102. ICRC.

[106] Independent High-Level Expert Group on Artificial Intelligence. 2020. *Ethics Guidelines for Trustworthy AI.* Technical Report. Council of Europe.

[107] Nora Osmani et al. 2020. The Complexity of Criminal Liability of AI Systems. *Masaryk University Journal of Law and Technology* 14, 1 (2020), 53–82.

[108] David Owens. 2017. *Trusting a Promise and Other Things.* Oxford University Press, Oxford, United Kingdom, 214–29.

[109] Hyanghee Park, Daehwan Ahn, Kartik Hosanagar, and Joonhwan Lee. 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 154, 15 pages. https://doi.org/10.1145/3411764.3445304

[110] Philip Pettit. 1995. The cunning of trust. *Philosophy & Public Affairs* 24, 3 (1995), 202–225.

[111] Azzurra Pini, Jer Hayes, Connor Upton, and Medb Corcoran. 2019. AI Inspired Recipes: Designing Computationally Creative Food Combos. In *CHI EA '19: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3290607.3312948

[112] Marianne Promberger and Jonathan Baron. 2006. Do patients trust computers? *Journal of Behavioral Decision Making* 19, 5 (2006), 455–468.

[113] Martin Ragot, Nicolas Martin, and Salomé Cojean. 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*. Association for Computing Machinery, New York, NY, USA, 1–10.

[114] Ilana Ritov and Jonathan Baron. 1992. Status-quo and omission biases. *Journal of Risk and Uncertainty* 5 (1992), 49–61.

[115] Kit T Rodolfa, Hemank Lamba, and Rayid Ghani. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nature Machine Intelligence* 3, 10 (2021), 896–904.

[116] Rocío Sánchez-Salmerón, José L Gómez-Urquiza, Luis Albendín-García, María Correa-Rodríguez, María Begoña Martos-Cabrera, Almudena Velando-Soriano, and Nora Suleiman-Martos. 2022. Machine learning methods applied to triage in emergency services: A systematic review. *International Emergency Nursing* 60 (2022), 101109.

[117] Filippo Santoni de Sio and Jeroen Van den Hoven. 2018. Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI* 5 (2018), 15.

[118] Anuschka Schmitt, Thiemo Wambsganss, Matthias Söllner, and Andreas Janson. 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)*, Vol. 1. ICIS, Austin, Texas, 1–17.

[119] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.

[120] Matthias Söllner, Axel Hoffmann, Holger Hoffmann, and Jan Marco Leimeister. 2012. How to use behavioral research insights on trust for HCI system design. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1703–1708.

[121] Mark Spranca, Elisa Minsk, and Jonathan Baron. 1991. Omission and commission in judgment and choice. *Journal of Experimental Social Psychology* 27, 1 (1991), 76–105. https://doi.org/10.1016/0022-1031(91)90011-T

[122] S Shyam Sundar. 2008. *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative, Chicago, IL, USA.

[123] S Shyam Sundar and Jinyoung Kim. 2019. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–9.

[124] S Shyam Sundar, Jinyoung Kim, Mary Beth Rosson, and Maria D Molina. 2020. Online privacy heuristics that predict information disclosure. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–12.

[125] Steven C Sutherland, Casper Harteveld, and Michael E Young. 2015. The role of environmental predictability and costs in relying on automation. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2535–2544.

[126] the Committee of Ministers of the Council of Europe. 2020. *Recommendation CM/Rec(2020)1 of the Committee of Ministers to member States on the human rights impacts of algorithmic systems*. Technical Report. Council of Europe.

[127] Thomas Theodoridis, Vassilios Solachidis, Kosmas Dimitropoulos, Lazaros Gymnopoulos, and Petros Daras. 2019. A survey on AI nutrition recommender systems. In *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. Association for Computing Machinery, New York, NY, United States, 540–546.

[128] Judith Jarvis Thomson. 1976. Killing, letting die, and the trolley problem. *The Monist* 59, 2 (1976), 204–217.

[129] Neil Thurman, Judith Moeller, Natali Helberger, and Damian Trilling. 2019. My friends, editors, algorithms, and I: Examining audience attitudes to news selection. *Digital Journalism* 7, 4 (2019), 447–469.

[130] Daniel W Tigard. 2021. Responsible AI and moral responsibility: a common appreciation. *AI and Ethics* 1, 2 (2021), 113–117.

[131] Peter M Todd and Gerd Gigerenzer. 2000. Précis of simple heuristics that make us smart. *Behavioral and brain sciences* 23, 5 (2000), 727–741.

[132] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. 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*. Association for Computing Machinery, New York, NY, USA, 77–87.

[133] Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. 2020. Implementations in machine ethics: a survey. *ACM Computing Surveys (CSUR)* 53, 6 (2020), 1–38.

[134] Amos Tversky and Daniel Kahneman. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* 5 (1992), 297–323.

[135] Kailas Vodrahalli, Tobias Gerstenberg, and James Zou. 2021. Do Humans Trust Advice More if it Comes from AI? An Analysis of Human-AI Interactions. *CoRR* abs/2107.07015 (2021), 1–34. arXiv:2107.07015 https://arxiv.org/abs/2107.07015

[136] Michael A Wallach, Nathan Kogan, and Daryl J Bem. 1964. Diffusion of responsibility and level of risk taking in groups. *The Journal of Abnormal and Social Psychology* 68, 3 (1964), 263.

[137] Dakuo Wang, Elizabeth Churchill, Pattie Maes, Xiangmin Fan, Ben Shneiderman, Yuanchun Shi, and Qianying Wang. 2020. From human-human collaboration to Human-AI collaboration: Designing AI systems that can work together with people. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[138] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. *Designing Theory-Driven User-Centric Explainable AI*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3290605.3300831

[139] Ori Weisel and Shaul Shalvi. 2015. The collaborative roots of corruption. *Proceedings of the National Academy of Sciences* 112, 34 (2015), 10651–10656.

[140] Anja Wölker and Thomas E Powell. 2021. Algorithms in the newsroom? News readers' perceived credibility and selection of automated journalism. *Journalism* 22, 1 (2021), 86–103.

[141] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. *Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[142] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. *Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[143] Karen Yeung. 2020. Recommendation of the council on artificial intelligence (oecd). *International Legal Materials* 59, 1 (2020), 27–34.

[144] Scot D. Yoder. 1998. The Nature of Ethical Expertise. *The Hastings Center Report* 28, 6 (1998), 11–19. http://www.jstor.org/stable/3528262

[145] Qiaoning Zhang, X Jessie Yang, and Lionel Peter Robert. 2020. Expectations and trust in automated vehicles. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–9.

[146] Yunfeng Zhang, Rachel KE Bellamy, Moninder Singh, and Q Vera Liao. 2020. Introduction to AI Fairness. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–4.