June 2022

Capable but Amoral?

Comparing AI and Human Expert Collaboration in Ethical Decision Making

Suzanne TOLMEIJER [a,1], Markus CHRISTEN [a], Serhiy KANDUL [a], Markus KNEER [a], and Abraham BERNSTEIN [a]

[a] *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.

Keywords. Ethical AI, Trust, Responsibility, Human-AI Collaboration

1. Introduction

In this extended abstract, we present work that is currently under review at the ACM CHI Conference on Human Factors in Computing Systems 2022.

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 [1] and co-creation of art [2] to HR decisions [3] and clinical decision making [4,5]. 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 [7]. However, another set of experiments has shown indications of users experiencing algorithmic aversion. For instance, people lose trust in

[1] Corresponding Author: Suzanne Tolmeijer; E-mail: tolmeijer@ifi.uzh.ch.

AI faster when it makes mistakes than when a human expert does [8]. 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 [9]. All of the mentioned factors that can trigger algorithmic aversion depend on the decision domain and task type the AI performing in [10].

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., [11,12]). 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 [13,14,15]. They are further divided over the question in virtue of what an action is to be assessed as right or wrong. Kant [16] famously placed strong emphasis on the agent's intentions, while consequentialists, such as Bentham [17] and Mill [18], 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 concerning the COMPAS recidivism algorithm [19]. It is mathematically impossible to adhere to all of people's different notions of fairness [20,21]. Instead, the transparency that algorithms offer for discrimination and bias in decision making highlight the trade-offs between different values [22]. While research on implementing ethics in AI has being ongoing, it has been in a scattered and relatively limited fashion [23].

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 [24]. 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 [11]. 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 [25]. 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 [26]. Generally, users assign more responsibility to parties that have more autonomy in decision making [27]. Different types of agency lead to different responsibility ascriptions, such as to the AI artifact, the designer, and the user of the system [28]. The assigned responsibility also depends on the role and autonomy the AI has [29].

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 [30]. 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 [31], 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 [32]. Chiang and Tin [33] 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 [34].

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., [35,36,37,38]), 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) [39]. 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 [40], 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 [41], 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

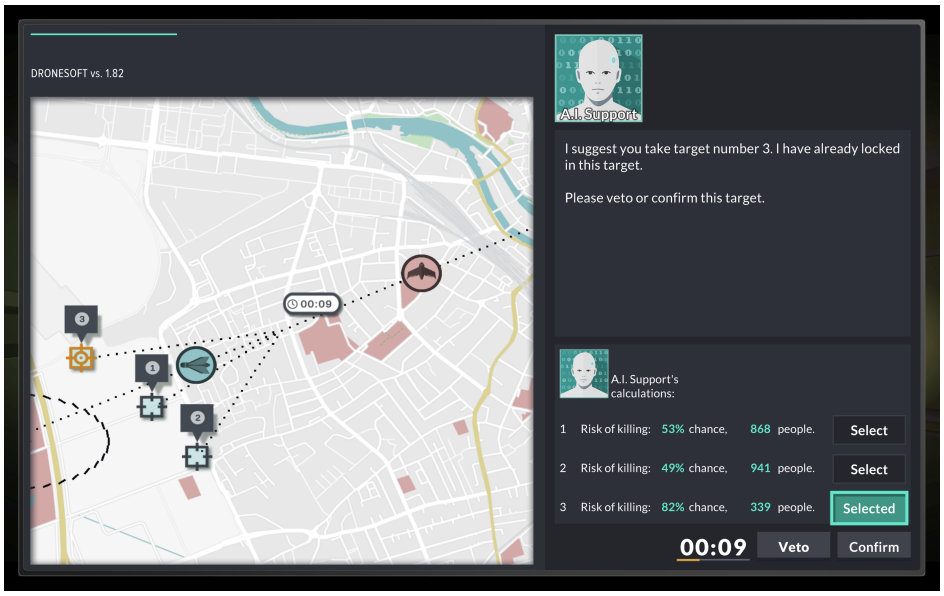

Figure 1. Screenshot of the software used to display dilemmas. The top-right shows the expert's avatar and advice, the bottom-right shows the different decision options and which advice the expert gave. The left depicts in interactive map on which the scenario unfolds.

on the expert's suggestion. The scenarios were presented using software developed by a game development company — a screenshot can bee found in Figure 1.

The described scenarios and software 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.

References

[1] Pini A, Hayes J, Upton C, Corcoran M. 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. CHI EA '19. New York, NY, USA: Association for Computing Machinery; 2019. p. 1–6. Available from: `https://doi.org/10.1145/3290607.3312948`.

[2] Li Z, Wang Y, Wang W, Greuter S, Mueller FF. Empowering a Creative City: Engage Citizens in Creating Street Art through Human-AI Collaboration. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. CHI EA '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 1–8. Available from: https://doi.org/10.1145/3334480.3382976.

[3] Park H, Ahn D, Hosanagar K, Lee J. Human-AI Interaction in Human Resource Management: Understanding Why Employees Resist Algorithmic Evaluation at Workplaces and How to Mitigate Burdens. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. New York, NY, USA: Association for Computing Machinery; 2021. Available from: https://doi.org/10.1145/3411764.3445304.

[4] Yang Q, Steinfeld A, Zimmerman J. In: Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. New York, NY, USA: Association for Computing Machinery; 2019. p. 1–11. Available from: https://doi.org/10.1145/3290605.3300468.

[5] Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Bermúdez i Badia S. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21. New York, NY, USA: Association for Computing Machinery; 2021. Available from: https://doi.org/10.1145/3411764.3445472.

[6] Araujo T, Helberger N, Kruikemeier S, De Vreese CH. In AI we trust? Perceptions about automated decision-making by artificial intelligence. AI & SOCIETY. 2020;35(3):611-23.

[7] Logg JM, Minson JA, Moore DA. Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes. 2019;151:90-103.

[8] Dietvorst BJ, Simmons JP, Massey C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General. 2015;144(1):114.

[9] Burton JW, Stein MK, Jensen TB. A systematic review of algorithm aversion in augmented decision making. Journal of Behavioral Decision Making. 2020;33(2):220-39.

[10] Castelo N, Bos MW, Lehmann DR. Task-dependent algorithm aversion. Journal of Marketing Research. 2019;56(5):809-25.

[11] Santoni de Sio F, Van den Hoven J. Meaningful human control over autonomous systems: A philosophical account. Frontiers in Robotics and AI. 2018;5:15.

[12] Fast E, Horvitz E. Long-term trends in the public perception of artificial intelligence. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31. Palo Alto, California USA: The AAAI Press; 2017. p. 963-9.

[13] Mackie J. Ethics: Inventing right and wrong. London, United Kingdom: Penguin UK; 1990.

[14] Harman G. Moral relativism defended. The Philosophical Review. 1975;84(1):3-22.

[15] Graham PA. In defense of objectivism about moral obligation. Ethics. 2010;121(1):88-115.

[16] Kant I. Groundwork of the metaphysics of morals. Cambridge, United Kingdom: Cambridge; 1785.

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

[18] Mill JS. 1998. Utilitarianism, edited with an introduction by Roger Crisp. New York, NY, USA: Oxford University Press; 1861.

[19] Khademi A, Honavar V. Algorithmic bias in recidivism prediction: A causal perspective (student abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34. Palo Alto, California USA: The AAAI Press; 2020. p. 13839-40.

[20] Loi M, Christen M. How to include ethics in machine learning research. ERCIM News. 2019;116(3):5.

[21] Kleinberg J, Mullainathan S, Raghavan M. Inherent Trade-Offs in the Fair Determination of Risk Scores. In: Papadimitriou CH, editor. 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). vol. 67 of Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik; 2017. p. 43:1-43:23.

[22] Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR. Discrimination in the Age of Algorithms. Journal of Legal Analysis. 2018;10:113-74.

[23] Tolmeijer S, Kneer M, Sarasua C, Christen M, Bernstein A. Implementations in machine ethics: a survey. ACM Computing Surveys (CSUR). 2020;53(6):1-38.

[24] Matthias A. The responsibility gap: Ascribing responsibility for the actions of learning automata. Ethics and information technology. 2004;6(3):175-83.

[25] of the Red Cross IR. Artificial intelligence and machine learning in armed conflict: A human-centred approach. ICRC; 2019. 102.

[26] Tigard DW. Responsible AI and moral responsibility: a common appreciation. AI and Ethics.

2021;1(2):113-7.
[27] Hong JW, Williams D. Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent. Computers in Human Behavior. 2019;100:79-84.
[28] Johnson DG, Verdicchio M. AI, agency and responsibility: the VW fraud case and beyond. Ai & Society. 2019;34(3):639-47.
[29] Lima G, Cha M. Descriptive AI Ethics: Collecting and Understanding the Public Opinion. In: Ethics in Design Workshop. vol. 1. New York, NY, USA: Association for Computing Machinery; 2020. p. 1-6.
[30] Nahavandi S. Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. IEEE Systems, Man, and Cybernetics Magazine. 2017;3(1):10-7.
[31] Methnani L, Aler Tubella A, Dignum V, Theodorou A. Let Me Take Over: Variable Autonomy for Meaningful Human Control. Frontiers in Artificial Intelligence. 2021;4:133.
[32] Lee JD, See KA. Trust in automation: Designing for appropriate reliance. Human factors. 2004;46(1):50-80.
[33] Chiang CW, Yin M. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In: 13th ACM Web Science Conference 2021. New York, NY, USA: Association for Computing Machinery; 2021. p. 120-9.
[34] Adnan N, Nordin SM, bin Bahruddin MA, Ali M. How trust can drive forward the user acceptance to the technology? In-vehicle technology for autonomous vehicle. Transportation research part A: policy and practice. 2018;118:819-36.
[35] Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, et al. The moral machine experiment. Nature. 2018;563(7729):59-64.
[36] Lin P. In: Why ethics matters for autonomous cars. Heidelberg, Germany: Springer, Berlin, Heidelberg; 2016. p. 69-85.
[37] Gogoll J, Müller JF. Autonomous cars: in favor of a mandatory ethics setting. Science and engineering ethics. 2017;23(3):681-700.
[38] Bonnefon JF, Shariff A, Rahwan I. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. Proceedings of the IEEE. 2019;107(3):502-4.
[39] Christen M, Narvaez D, Zenk JD, Villano M, Crowell CR, Moore DR. Trolley dilemma in the sky: Context matters when civilians and cadets make remotely piloted aircraft decisions. PLoS one. 2021;16(3):e0247273.
[40] Thomson JJ. Killing, letting die, and the trolley problem. The Monist. 1976;59(2):204-17.
[41] Bauman CW, McGraw AP, Bartels DM, Warren C. Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. Social and Personality Psychology Compass. 2014;8(9):536-54.