Overtrust in Algorithms: An online behavioral study on trust and reliance in AI advice

Phillipp Schreck, Martin Luther University of Halle-Wittenberg (Germany)
Artur Klingbell, Martin Luther University of Halle-Wittenberg (Germany)
Cassandra Grüzner, Martin Luther University of Halle-Wittenberg (Germany)
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: human machine interaction, decision making, behavioral experiment, trustworthy AI, algorithm appreciation

While technical innovations in artificial intelligence (AI) are evolving at a rapid pace, one practical area that AI has already heavily impacted is decision-making. Where humans previously had to rely on their own experience from prior situations, empirically developed heuristics or historical benchmarks of potential indicators, nowadays many complex decisions are supported by AI systems. Benefits of adopting AI can include more informed decisions, as the AI can process significant amounts of information effectively, or fairer results due to the removal of human flaws such as cognitive bias (Agarwal et al., 2018; Bozdag, 2013; Savulescu & Maslen, 2015). However, relying blindly on these technologies also harbors potential dangers: AI systems can introduce drawbacks such as immoral results due to algorithmic bias (Diakopoulos, 2015). To understand when and why individuals may overtrust the recommendation of an AI, we are conducting a behavioral online study, investigating under which circumstances agents follow AI advice - despite the advice going counter to their own decision-making intuitions.

Since *Computers Are Social Actors* (CASA) research upholds that humans react socially to machines (Nass & Moon, 2000), and trust is one of the major factors influencing the human-machine interaction (Lee & Moray, 1992), to secure a productive and sustainable integration of technology in the future, it is crucial to examine how trust in machines influences their use – both in positive and in negative ways. Trust is needed to foster successful utilization of new technology, since a lack thereof might result in systems not being handled to their full potential, reducing the overall rate of adoption or ultimately leading to disuse of the technology. Blind trust in a system, whose capabilities do not warrant that trust, however can also have detrimental effects, as misuse of technology might lead to undesired consequences both from an economic as well as an ethical perspective. Instead, calibrated trust should be aimed for, meaning trust that matches a system's capabilities, hence leading to appropriate use (Lee & See, 2004)

Various social and psychological mechanics can drive people to act in a way that they know to be wrong. In their classic experiment on obedience to authority Milgram and Gudehus (1978) studied how people follow orders of authority figures against better judgment even consciously performing unethical actions as long as they do not own the responsibility. Similarly, Asch's (1951) study of independence and conformity examined how people behave under social pressures. While some participants remained independent and resisted group pressure, others changed their behavior under demanding conditions and behaved in a way that conformed to expectations against

their own better judgment. Furthermore, Bandura (1986) proposed that people can engage in immoral behavior without believing they are doing anything wrong by using a cognitive process called moral disengagement. Trusting a machine over the appropriate amount might similarly allow for agents to make wrong decisions against their better knowledge. Leicht-Deobald et al. (2019) propose that reliance on algorithmic decision-making could lead to blind trust in rules, where compliance trumps personal integrity. Thus, overtrust meaning the extension of unjustified trust and potentially resulting overreliance in following AI-generated recommendations might stem from multiple sources but could lead to potentially dangerous consequences ranging from slightly overcredulous behavior to blind obedience to the system.

Recent experimental insights seem to support these doubts, as studies are questioning the role of the human-in-the-loop and the necessity of algorithmic trustworthiness for algorithms to be trusted. In theory, Explainable AI (XAI) should reduce overreliance on AI by enabling humans to assess AI recommendations and synthesize them with their own judgment. However, several studies indicate that explanations, while successful in increasing trust, frequently also lead to overreliance, as users followed wrong AI suggestions despite provided explanations and against better knowledge (e.g., Bussone et al., 2015; Jacobs et al., 2021; Lai & Tan, 2019). Bansal et al. (2021) suggest that supplying explanations might even increase the chance that humans accept AI advice regardless of its accuracy. Buçinca et al. (2021) argue that to mitigate overreliance, providing explanations alone is insufficient. Regarding trustworthiness Krügel et al. (2022, 2023) have examined the adherence to AI-advisors in ethical decision-making situations, and are suggesting that overtrust in AI is more prevalent than distrust despite low trustworthiness.

Still, behavioral research on whether human agents tend to follow algorithmic suggestions has resulted in conflictive results, such as seen in the renowned studies on algorithm aversion (Dietvorst et al., 2015; Dzindolet et al., 2002) versus algorithm appreciation (Logg et al., 2019). While these two research streams seem to provide contradicting conclusions, the question arises which factors lead to higher or lower trust. Some scholars claim that the

differences can be explained with framing. How people actually behave seems to heavily depend on the setup of the individual study. Specifically, manipulating only the description of the human and the algorithmic agent can yield drastically opposing results regarding adherence to the algorithm (Hou & Jung, 2021). Some scholars study the relationship between trust and system performance measured in terms of stated accuracy or observed accuracy (Kennedy et al., 2022; Yin et al., 2019; Yu et al., 2016; Yu et al., 2017). Others examine how a model's interpretability influences trust (Poursabzi-Sangdeh et al., 2021; Ribeiro et al., 2016) or how confidence scores and explanations effect trust calibration (Zhang et al., 2020).

Nevertheless, these studies were mostly focused on specific scenarios and limited tasks. Hence, studies on overtrust in ethically relevant recommendations by algorithms are still lacking a generalizable experimental setup that is more context-independent. Additionally, as evidence on overtrust is conflicting, further research is required to understand which factors promote it and which contrarily foster distrust in algorithms.

In our study, we aim to address both the issue of high context-sensitivity in studies regarding algorithm aversion or appreciation, as well as the issue of understanding which factors may lead to overtrust and overreliance on algorithmic advice. To address the first issue, we are following behavioral economic conventions and are using a decontextualized game to study our participants' behavior. By basing our inquiry on a modified version of the context-free trust game by Berg et al. (1995), we can infer implications for various ethically-relevant applications such as recruiting decisions or loan approvals while eliminating the need to use elaborate context descriptions or highly specific tasks. Therefore we conduct behavioral online experiments in order to assess overtrust and overreliance during ethically-relevant decision-making situations. Participants play several rounds of an incentivized, interactive, repeated ethical dilemma game, in which pairs (consisting of one trustee and one trustor) are randomly rematched every round. Every round the trustee decides whether to keep their endowment or to invest it by allocating their complete funds to the trustor, which in turn is tripled. Afterwards the trustor decides whether to keep the amount for himself or to return half to the trustee. In a predetermined round the trustee receives advice on how to act based on the decision history of the newly matched trustor. This recommendation is presented as stemming either from an expert or from an AI, in this case a

generative pre-trained transformer, that was instructed to give recommendations based on the previous decision history.

To address the second issue, of understanding which factors may lead to overtrust and overreliance, we are focussing on whether and when agents are willing to follow advice that is counterintuitive. While in a lot of research, it is easy to define what can be considered the appropriate amount of trust, and thus, if an individual is overtrusting the AI, the real world is rarely as clear. Often, it is impossible to say upfront if an individual is overtrusting the algorithm they are encountering, either because it is not easily possible to trace how the decisions have been made within the AI, or because there simply are no clear true or false answers to many complicated problems. We mirror this in our experiment, as there is no reliable, failsafe way to predict the trustor's actions until they already take place. The willingness to follow counterintuitive advice therefore acts as an indicator of overtrust, as agents are trusting the advice above their own good judgment. Seeing as the advice is only given a single time, players do not have any experiences that may lead them to believe that provided advice is more reliable than their own decisions. Thus, we should expect a rejection of the advice given in at least the most counterintuitive situations. To judge those, we are both looking at self-reported intuitions of the participants, as well as deriving where most players' intuitions lay given the same playing history by looking at a control group that plays without being given any advice.

Subsequently we are examining which conditions foster subjects to demonstrate overtrust in counter-intuitive AI recommendations and which result in subjects actively reversing the algorithm's advice. For this purpose, we plan to manipulate various factors such as available information about the algorithm, its perceived competence, reliability, decision complexity, decision amount and restrictions such as budget or time limits. Further, we are also interested in which personal factors may lead to overtrust, by looking at risk aversion, technological affinity and trusting personality traits. The goal of our study is to derive more generalizable information about which factors influence potentially harmful overtrust in AI and how this can be mitigated.

REFERENCES

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification. International Conference on Machine Learning.

AI HLEG. (2019). Ethics Guidelines For Trustworthy AI. High-Level Expert Group on Artificial Intelligence (AI HLEG) set up by the European Commission.

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. Organizational influence processes, 58, 295-303.

Bandura, A. (1986). Social foundations of thought and action. Englewood Cliffs, NJ, 1986(23-28).

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. Calif. L. Rev., 104, 671.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. Games and Economic Behavior, 10(1), 122-142.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. Ethics and information technology, 15(3), 209-227.

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-21.

Bussone, A., Stumpf, S., & O'Sullivan, D. (2015). The role of explanations on trust and reliance in clinical decision support systems. 2015 international conference on healthcare informatics.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. Digital journalism, 3(3), 398-415.

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors, 44(1), 79-94.

Hou, Y. T.-Y., & Jung, M. F. (2021). Who is the expert? Reconciling algorithm aversion and algorithm appreciation in AI-supported decision making. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-25.

Jacobs, M., Pradier, M. F., McCoy Jr, T. H., Perlis, R. H., Doshi-Velez, F., & Gajos, K. Z. (2021). How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. Translational psychiatry, 11(1), 108.

Kennedy, R. P., Waggoner, P. D., & Ward, M. M. (2022). Trust in public policy algorithms. The Journal of Politics, 84(2), 1132-1148.

Koulish, R. (2016). Using risk to assess the legal violence of mandatory detention. Laws, 5(3), 30.

Krämer, N. C., Rosenthal-von der Pütten, A. M., & Hoffmann, L. (2015). Social effects of virtual and robot companions. The handbook of the psychology of communication technology, 32, 137-137.

Krügel, S., Ostermaier, A., & Uhl, M. (2022). Zombies in the Loop? Humans Trust Untrustworthy AI-Advisors for Ethical Decisions. Philosophy & Technology, 35(1), 1-37. Krügel, S., Ostermaier, A., & Uhl, M. (2023). Algorithms as partners in crime: A lesson in ethics by design. Computers in Human Behavior, 138, 107483.

Lai, V., & Tan, C. (2019). On human predictions with explanations and predictions of machine learning models: A case study on deception detection. Proceedings of the conference on fairness, accountability, and transparency.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50-80.

Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243-1270.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I., & Kasper, G. (2019). The challenges of algorithm-based HR decision-making for personal integrity. Journal of Business Ethics, 160(2), 377-392.

Lewicki, R. J., & Wiethoff, C. (2000). Trust, trust development, and trust repair. The handbook of conflict resolution: Theory and practice, 1(1), 86-107.

Logg, J. M., Minson, J. A., & Moore, D. A. (2019). Algorithm appreciation: People prefer algorithmic to human judgment. Organizational Behavior and Human Decision Processes, 151, 90-103.

Milgram, S., & Gudehus, C. (1978). Obedience to authority. In: Ziff-Davis Publishing Company New York, NY, USA.

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. Journal of social issues, 56(1), 81-103.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., & Wallach, H. (2021). Manipulating and measuring model interpretability. Proceedings of the 2021 CHI conference on human factors in computing systems.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining.

Savulescu, J., & Maslen, H. (2015). Moral enhancement and artificial intelligence: moral AI? In Beyond artificial intelligence (pp. 79-95). Springer.

Yin, M., Wortman Vaughan, J., & Wallach, H. (2019). Understanding the effect of accuracy on trust in machine learning models. Proceedings of the 2019 chi conference on human factors in computing systems.

Yu, K., Berkovsky, S., Conway, D., Taib, R., Zhou, J., & Chen, F. (2016). Trust and reliance based on system accuracy. Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. Proceedings of the 22nd international conference on intelligent user interfaces.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020). Effect of confidence and explanation on

accuracy and trust calibration in Al-assisted decision making. Proceedings of the 2020 conference on fairness, accountability, and transparency.