Overtrust in algorithms – An online behavioral study

Philipp Schreck

Martin-Luther-University Halle-Wittenberg, philipp.schreck@wiwi.uni-halle.de

Artur Klingbeil

Martin-Luther-University Halle-Wittenberg, artur.klingbeil@wiwi.uni-halle.de

Cassandra Grützner

Martin-Luther-University Halle-Wittenberg, cassandra.gruetzner@wiwi.uni-halle.de

Inappropriate reliance on technology can result in misuse of systems potentially leading to unethical decisions due to algorithmic bias, unprofitable outcomes for the affected parties or even safety hazards when operators fail to assume control in a situation where machines err. While evidence hints on overtrust in algorithms, further research is required to understand which factors promote it and which contrarily foster distrust in algorithms. However, most of the research has focused on specific scenarios and limited tasks. Hence, studies on ethically relevant recommendations by algorithms are lacking a generalizable experimental setup that is context-independent of general factors that may lead to overtrust.

To address this, we utilize methodologies from behavioral economics to conduct online experiments. We assess under which conditions subjects demonstrate overtrust in counter-intuitive AI recommendations made by a generative pre-trained transformer during decision-making situations and when subjects actively reverse the algorithm's recommendation. For this purpose, we plan to manipulate factors such as available information about the algorithm, its perceived competence, reliability, decision complexity, decision amount and restrictions such as budget or time limits. The goal is to derive more generalizable principles from a behavioral perspective about how to design ethical AI that does not foster potentially harmful overtrust in machines.

CCS CONCEPTS • Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI • Social and professional topics~Professional topics~Computing and business~Computer supported cooperative work • Computing methodologies~Artificial intelligence~Natural language processing • Computing methodologies~Artificial intelligence~Philosophical/theoretical foundations of artificial intelligence

Additional Keywords and Phrases: Trust and reliance, human computer interaction, decision making, behavioral experiment, algorithm appreciation, trustworthy AI

1 EXTENDED ABSTRACT

In our talk, we are presenting our work in progress of an online experiment based on behavioral economic methodology, investigating which factors lead people to experience overtrust in Artificial Intelligence (AI). We are especially interested in cases in which counter-intuitive AI generated recommendations are accepted, and when participants actively reverse and correct a machine decision.

One central field in the behavioral interaction between humans is the study of trust and trustworthiness, examining how trust is formed initially, developed during cooperation, broken by violations, and restored with trust repair attempts (e.g., Lewicki & Wiethoff, 2000). Research from the Computers Are Social Actors (CASA) paradigm has shown that people feel affiliated to their team members regardless of them being humans or computers (Nass et al., 1996), that people collaborate similarly with human and human-like virtual companions (Krämer et al., 2015), and that people interacting with computers can perform various social behaviors (e.g. politeness) although they are aware that the machine does not feel any emotions (Nass et al., 1994; Reeves & Nass, 1996). Since 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. Nevertheless, too much trust in a system, whose capabilities do not warrant that trust, can also have detrimental effects, as misuse of technology might lead to undesired consequences both from an economic as well as an ethical perspective. Overtrust in machines might range from slightly overcredulous behavior to blind obedience to the system, potential dangers of which might include a qualified applicant being rejected by a recruiter due to algorithmic discrimination, a medical doctor failing to diagnose an illness that was not highlighted due to incomplete training data in the detection device or an accident caused by a driver refusing to assume manual control of his malfunctioning autonomous vehicle.

While technical innovations in AI are ongoing, one practical area that AI has already heavily impacted is decision-making. Where previously humans 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. For example, algorithmic decisions aids are used in HR departments for hiring/recruiting, in insurance companies to classify risk, or in banks to assess credit rating. In these situations, AI is frequently employed to provide additional information or to create decision recommendations. 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).

Despite its various advantages delegating decision competence to AI introduces other risks, hence implementing a human-in-the-loop or related concepts like human-in-command have been frequently suggested to retain the ultimate decision authority to human reasoning. These calls have been implemented in various (inter-)national initiatives e.g., in its “ethics guidelines for trustworthy AI” the European Union specifies that human agency and oversight are critical requirements to achieving trustworthy AI, such as the ability to decide “not to use an AI system” or “to override a decision made by a system” (AI HLEG, 2019, p.16). These calls for human oversight together with other factors concerning the algorithm such as transparency, accuracy and accessibility oftentimes constitute the requirements for trustworthy AI.

However, there are reasons to doubt both, the effectiveness of the human monitoring function as well as the necessity of trustworthiness as a prerequisite to (unjustified) trust. 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 to do anything wrong by using a cognitive process called moral disengagement. Since various social and psychological mechanics can lead people to act in a way that they know to be wrong, overtrust in machines might similarly result in people making wrong decisions against better knowledge. Leicht-Deobald et al. (2019) propose that reliance on the algorithmic decision-making could lead to blind trust in rules, where compliance trumps personal integrity. Thus, overtrust meaning the extension of unjustified trust and following in AI-generated recommendations might stem from multiple sources but is potentially dangerous.

While the use of AI can benefit the decision-making process, it can also introduce drawbacks, such as immoral results due to algorithmic bias (Diakopoulos, 2015). Well-known instances of immoral actions due to biased AI decisions include discrimination against marginalized groups (Barocas & Selbst, 2016), and unnecessarily high detention rates of immigrants in the US (Koulish, 2016). In the hope to resolve these issues by introducing a human-in-the-loop, the situation might even have worsened. If humans overtrust the AI recommendation, they may reinforce existing bias and misjudgments, while simultaneously believing the system to generate better outcomes due to the human oversight.

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. E.g., Krügel et al. (2022, 2023) have examined the adherence to AI-advisors in ethical decision-making situations, and are suggesting that regardless of its trustworthiness, overtrust in AI is more prevalent than distrust. 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, while evidence hints on overtrust, further research is required to understand which factors promote it and which contrarily foster distrust in algorithms.

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) versus algorithm appreciation (Logg et al., 2019). While these two research streams seem to provide contradicting conclusions, 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).

In our study, we aim to address both the issue of understanding which factors may lead to overtrust, as well as the issue of high context-sensitivity in studies regarding algorithm aversion or appreciation. Therefore, we conduct behavioral online experiments in order to assess overtrust during ethically-relevant decision-making situations. In an interactive repeated ethical dilemma game, participants receive advice on how to decide in a specific round. This recommendation stems 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. We are examining under which

conditions subjects demonstrate overtrust in counter-intuitive AI recommendations and when subjects actively reverse 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. The goal of our study is to derive more generalizable information about which factors influence potentially harmful overtrust in AI.

SELECTION OF 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).

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

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

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': the importance of trust repair in human–machine interaction. Ergonomics, 61(10), 1409-1427.

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.

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.

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.

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.

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? International Journal of Human-Computer Studies, 45(6), 669-678.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. Proceedings of the SIGCHI conference on Human factors in computing systems,

Reeves, B., & Nass, C. (1996). The media equation: How people treat computers, television, and new media like real people. Cambridge university press Cambridge, UK.

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