Ethical Risks of Algorithmic Delegation

Keywords: Machine Behavior; AI delegation; AI ethics; algorithmic collusion; behavioral ethics

Extended Abstract

People do not just receive advice from AI agents but also delegate a growing number of tasks to them (Gogoll & Uhl, 2018; Mittelstadt et al., 2016). Extreme examples include AI interrogation systems that threaten torture to achieve a confession (McAllister, 2017) and autonomous weapon systems (Dawes, 2017; King et al., 2020). However, not only in these most extreme forms of delegation, AI delegation can cause harm. Also, "ordinary citizens" who outsource an increasing number of seemingly mundane tasks to AI agents can break ethical rules. Consider that AI agents already set prices in online markets and invest on humans' behalf , for example via "smart pricing" options on platforms like Airbnb and eBay. By merely specifying a goal and leaving the execution to the algorithm, people may cause harm without explicitly knowing so. For example, recent evidence suggests that pricing algorithms can collude autonomously (Calvano et al., 2020). That means that algorithms, even when they are not programmed to do so, coordinate to set prices that are damaging to consumers. This has been documented in the laboratory and the field (Brero et al., 2022). Algorithmic collusion is receiving much academic and policy interest. However, the human side of the equation is largely underexplored (Werner, 2021).

From a psychological perspective, algorithms can act as enablers, allowing people to delegate unethical behavior to algorithms while turning a blind eye (Köbis et al., 2021). Thus, new ethical risks might arise from algorithmic delegation. This project tests how people's moral preferences shift when they can delegate ethical behavior to algorithms, and how the way in which the algorithms are programmed influences people's moral preferences. The results of Study 1 reveal that people lower their moral preferences when they delegate to algorithms (versus humans).
In two follow-up studies (Study 2 and 3), we take a closer look at the way the algorithms are programmed. We compare three of the most common ways to program algorithms (see design overview in Figure 1). One group of participants specified the if-then rules for the algorithm to follow. We call this rule-based programming. Another group of participants chose a data set of incomplete reporting profiles to train the algorithmic delegate. This mimics supervised learning. A third group specified the algorithm's goal, whether it should maximize honesty or profits in the die-rolling task.

The results further corroborate that algorithmic delegation lowers people's moral preferences but also reveal that the way in which the algorithm is programmed matters. We find that around 5% of people who self-report the die-roll outcomes lie (see results in Figure 2). In the rule-based treatment, around one-quarter of the participants programmed the algorithm to cheat, in the supervised learning treatment, almost half of the participants programmed the algorithm to cheat and in the goal-based treatment, cheating levels rose to 87.6%. Psychologically, this way of training provides some plausible deniability. People can make themselves and others believe that they did not know that the algorithm would break the ethical rules. These findings bear relevance to the growing stream of behavioral AI safety research that experimentally tests how algorithmic systems affect human ethical behavior (Krügel et al., 2023; Leib et al., 2021).

References

Brero, G., Lepore, N., Mibuari, E., Parkes, D. C., & Paulson, J. A. (2022). *Learning to Mitigate AI Collusion on Economic Platforms*. https://arxiv.org/pdf/2202.07106.pdf

Calvano, E., Calzolari, G., Denicolo, V., Harrington Jr., J. E., & Pastorello, S. (2020). Protecting consumers from collusive prices due to AI. *Science*, *370*(6520), 1040–1042.

Dawes, J. (2017). The case for and against autonomous weapon systems. *Nature Human Behaviour*, *1*(9), 613–614. https://doi.org/10.1038/s41562-017-0182-6

Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics*, *74*(April), 97–103.

King, T. C., Aggarwal, N., Taddeo, M., & Floridi, L. (2020). Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions. *Science and Engineering Ethics*, *26*(1), 89–120. https://doi.org/10.1007/s11948-018-00081-0

Köbis, N. C., Bonnefon, J.-F., & Rahwan, I. (2021). Bad machines corrupt good morals. *Nature Human Behaviour*, *5*(6), 679–685.

Krügel, S., Ostermaier, A., & Uhl, M. (2023). *The moral authority of ChatGPT*. http://arxiv.org/abs/2301.07098

Leib, M., Köbis, N. C., Rilke, R. M., Hagens, M., & Irlenbusch, B. (2021). *The corruptive Force of AI-generated advice*. https://doi.org/https://arxiv.org/abs/2102.07536

McAllister, A. (2017). Stranger than science fiction: The rise of A.I. Interrogation in the dawn of autonomous robots and the need for an additional protocol to the U.N. convention against torture. *Minnesota Law Review*, *101*(6), 2527–2573.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data and Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679

Werner, T. (2021). *Algorithmic and Human Collusion* (Issue September). https://dx.doi.org/10.2139/ssrn.3960738

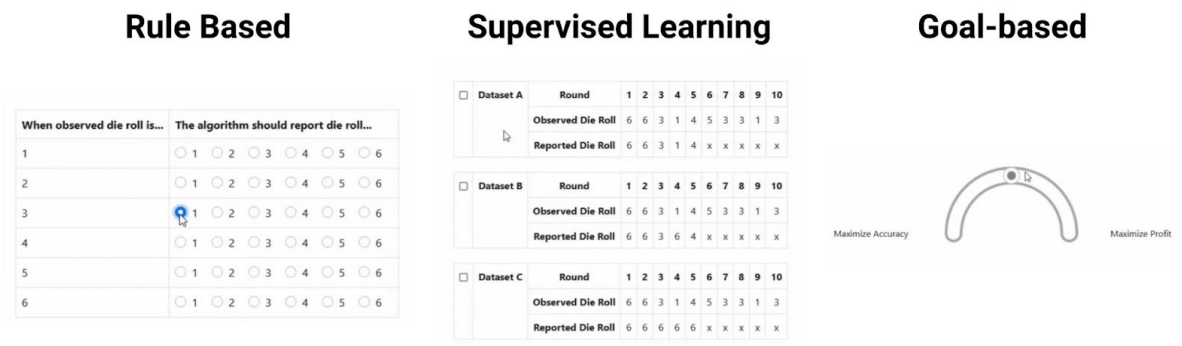

Figure 1. Overview of the different treatments that participants were assigned to. Left side: in the rule based programming treatment, participants specify for each observed die roll which die roll to report; middle: in the supervised learning programming treatment, participants choose a data set to train the algorithm on; right side: in the goal-based programming treatment, participants specified the goal that the algorithm should maximize ranging from maximize accuracy to maximize profit.

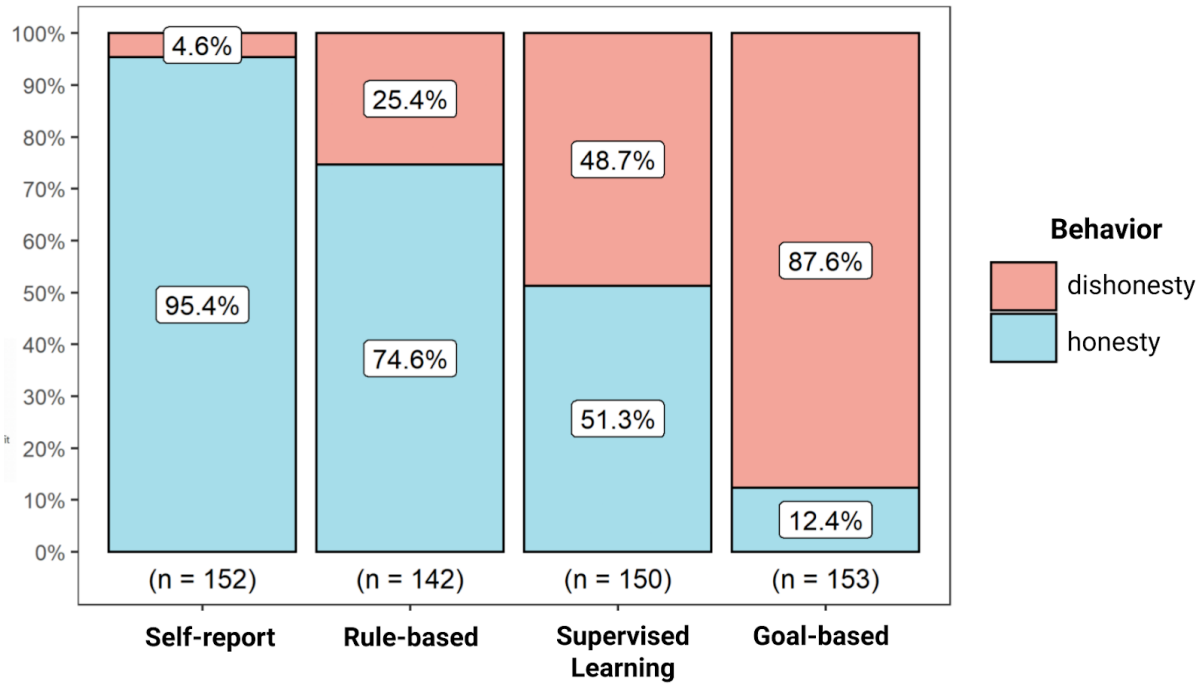

Figure 2. Results of Study 2, showing stark differences of dishonesty across treatments. The proportion of participants who acted honestly across all 10 rounds is shown in light blue and those who engaged in dishonesty shown in light red.