Framing, not transparency, reduces cheating in algorithmic delegation

Neele Engelmann
engelmann@mpib-berlin.mpg.de

Lara Kirfel
kirfel@mpib-berlin.mpg.de

Anne-Marie Nussberger
nussberger@mpib-berlin.mpg.de

Raluca Rilla
rilla@mpib-berlin.mpg.de

Iyad Rahwan
rahwan@mpib-berlin.mpg.de

Center for Humans and Machines, Max Planck Institute for Human Development, Berlin

Abstract

Recent evidence suggests that delegating tasks to machines can facilitate unethical behavior, but the psychological mechanisms driving this effect are not yet well understood. This study investigates whether two interventions can mitigate cheating in an algorithmic honesty game: transparency (information about which user input causes which algorithm behavior) and framing (natural language cues about the moral valence of behavior). In a 2×2 experimental design, we find that transparency does not reduce dishonest behavior, despite participants actively engaging with and understanding the provided information. Conversely, framing — replacing neutral labels like "maximize profit" with ethically charged terms like "maximize cheating" — substantially reduces dishonesty. These findings suggest that curbing misuse of AI requires confronting users with its moral implications, not just explaining the mechanics.

Keywords: human-AI interaction, AI ethics, moral psychology, causal reasoning

Introduction

With the rise of Artificial Intelligence (AI) and other advanced technologies, we are increasingly able to delegate tasks that once required our direct involvement (Candrian & Scherer, 2022). AI writing assistants streamline the creation of professional emails and persuasive documents, automated trading platforms execute rapid financial decisions to capitalize on market fluctuations, and pricing algorithms on platforms like Airbnb help hosts maximize earnings by adjusting rates to demand. These applications allow users to save time, optimize strategies, and unlock financial or strategic advantages that would be difficult to achieve manually. However, the behavior of algorithmic delegates can create significant risks or costs for others. AI-generated writing may mislead or manipulate recipients (Illia, Colleoni, & Zyglidopoulos, 2023), automated trading can exacerbate market volatility (International Monetary Fund, 2024), and dynamic pricing algorithms can inflate prices (MacKay & Weinstein, 2022). These risks may often not be foreseen by the users of such systems. But even when risks are anticipated, the mere act of delegating may allow users to turn a blind eye to potential externalities and still reap the rewards of their (algorithmic) delegate's activities (Bartling & Fischbacher, 2012). For instance, the average Airbnb host is unlikely to assemble other hosts in the neighborhood and actively conspire to inflate rent. But they may readily use a pricing algorithm which does just that, maybe even after coming across the occasional newspaper article that highlights the risk of algorithmic collusion [1]. Indeed, past work indicates that AI systems can sometimes absorb human responsibility for bad outcomes in joint action settings, highlighting the risk for strategic scapegoating of AI's in such contexts (Hohenstein & Jung, 2020; Parlangeli, Curro', Palmitesta, & Guidi, 2024; Feier, Gogoll, & Uhl, 2021; Shank, DeSanti, & Maninger, 2019). Turning a blind eye to negative externalities may become particularly easy for users when the inner workings of algorithms are opaque, potentially by facilitating moral disengagement (Bandura, 2002; Moore, 2015; Newman, Le, North-Samardzic, & Cohen, 2020).

Initial evidence for ethical risks of delegating to AI Köbis et al. (2024) recently provided empirical evidence that delegating to algorithms can facilitate unethical behavior. They used variations of the die-rolling game (Fischbacher & Föllmi-Heusi, 2013) to study how different delegation mechanisms influence user behavior. In the die-rolling game, participants privately observe a series of die rolls and are tasked with reporting the outcomes. Higher die-rolls result in higher bonus payments (e.g., 1 cent per pip), creating an incentive to cheat by inflating the reported outcomes. In Köbis et al. (2024)'s modification of this game (Study 2), participants could either report the die-roll outcomes themselves or delegate this task to an algorithmic agent. The algorithmic delegate could be instructed in one of three ways, each representing a different level of user involvement and abstraction. In the rule-based condition, participants explicitly programmed a fixed mapping between observed and reported outcomes for all possible die rolls, requiring direct and precise instructions. In the supervised learning condition, participants se-

[1] such as: https://www.theatlantic.com/ideas/archive/2024/08/ai-price-algorithms-realpage/679405/

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

lected one of three datasets to train the algorithm: one dataset modeled fully honest reporting, one represented partial dishonesty, and one showed maximal over-reporting. Finally, in the goal-based condition, participants adjusted a slider between abstract goals such as "maximize accuracy" and "maximize profit," without being shown how specific slider settings would influence the algorithm's behavior (see Figure 1A). Their results showed that delegation increased over-reporting, with dishonesty escalating as the delegation mechanism became more abstract or indirect: from 24% dishonesty in the rule-based condition to 52% in supervised learning, and 83% in the goal-based condition.

Psychological mechanisms and prevention What is it about opaque algorithmic delegates that seems to make requests for cheating so easy and appealing, and how can we mitigate this risk? Part of the explanation could be that the causal links between user input, algorithm behavior, and outcomes are obscured in these cases: It is unclear what exactly an algorithm will do when prompted by the user, and how harmful the side effects could be. Converging evidence from research in causal and moral reasoning (Lagnado & Gerstenberg, 2017; Engelmann & Waldmann, 2022; Waldmann, Wiegmann, & Nagel, 2017; Paharia, Kassam, Greene, & Bazerman, 2009; Royzman & Baron, 2002; Ziano et al., 2021; Weiss & Forstmann, 2024; Cushman, 2008) as well as behavioral economics (Oexl & Grossman, 2013; Hamman, Loewenstein, & Weber, 2010; Hill, 2015; Bartling & Fischbacher, 2012), indicates that with weak, indirect, or ambiguous causal connections between people's actions and harmful or unfair outcomes, perceived responsibility as well as blame and punishment are reliably diminished. In Köbis et al. (2024)'s "goal-based" condition, participants steered the algorithmic delegate's behavior by selecting one of seven settings between "maximize accuracy" and "maximize profit" on an otherwise unlabeled sliding scale (see Figure 1A). While participants could have made some reasonable assumptions about the algorithm's tendencies at the extremes, the concrete behavior and the exact amount of profit that could be expected at any given setting were never made explicit. Thus, uncertainty about both how much of an unfair bonus could be attained as well as about how exactly the algorithm would produce it may have enabled people to shoot for high profits while avoiding to see themselves as acting unethically in Köbis et al. (2024)'s "goal-based" delegation condition.

Transparency as a remedy? These considerations align with the now commonplace calls for transparency in AI deployment, both for deployers themselves and for potential auditors of systems. Transparency both about the presence of AI systems and about their inner workings is generally taken to be a prerequisite for trust, accountability and responsible use (Cheong, 2024; Akhtar, Kumar, & Nayyar, 2024; Lepri, Oliver, Letouzé, Pentland, & Vinck, 2018). Transparency for deployers is featured prominently in the EU AI act, specifically it demands that high-risk AI systems be "designed and developed in such a way as to ensure that their operation is sufficiently transparent to enable deployers to interpret a system's output and use it appropriately" (EU AI Act, Article 13). Regarding the algorithmic die-rolling game as a model of risky AI use, one way to improve transparency would be to render the causal links between user input, algorithm behavior, and outcomes explicit. Understanding that and how one's action produces negative consequences might make it much harder to instruct the algorithm to maximize profit while still seeing oneself as acting ethically. Hence, the first goal of this paper is testing the prediction that providing transparency about the causal relations between user input, algorithm behavior, and outcomes deters requests for over-reporting in the die-rolling game (**Hypothesis 1**). While the results obtained by Köbis et al. (2024) provide a first indication that transparency may matter, the modes of instructing the algorithm differ too strongly between conditions to draw any firm conclusions about the psychological mechanisms at play (programming a rule with six conditions vs. picking one of three sets of training data vs. selecting a setting on an opaque sliding scale). Thus, we are going to vary transparency while keeping the mode of instructing the algorithm constant, using the condition that produced the highest cheating rates in previous studies as our point of departure (selecting a setting on an opaque slider with labeled endpoints).

Transparency about what a system does on the mechanical level, however, may only be one side of the coin. Arguably, users also need to understand the ethical significance or meaning of an AI's behavior in order to deploy it responsibly (Ananny & Crawford, 2018; Kolkman, 2022; Felzmann, Fosch-Villaronga, Lutz, & Tamò-Larrieux, 2020; Andrada, Clowes, & Smart, 2023; Felzmann, Villaronga, Lutz, & Tamò-Larrieux, 2019). Besides obscuring the causal links between user input and algorithm behavior, Köbis et al. (2024)'s most opaque delegation condition is also the only one giving a natural language description to the system's behavior, notably a rather positive or at least neutral one - after all, what is wrong with "maximizing profit" in a paid online experiment? We suspect that natural language cues may be an additional factor that influences how users perceive and interact with AI systems. When the language used to describe an AI's functionality aligns with socially or economically desirable goals, it may encourage users to overlook or justify externalities. Conversely, more morally charged language, such as "maximize cheating," could alter users' perception of the ethical implications of their choices, and might therefore curb requests for over-reporting in the die-rolling game (**Hypothesis 2**). Testing this prediction is the second goal of this paper. Finally, it's possible that the effects of transparency and framing interact. For instance, natural language cues might more strongly affect people's behavior when they have little knowledge on the mechanical level about the system they are interacting with. While we don't predict any specific pattern here, we will test for possible interactions as well.

Experiment

We designed an incentivized online experiment to test the effects of transparency and framing on requests for over-reporting in a modified die-rolling game. The experiment was implemented in oTree (Chen, Schonger, & Wickens, 2016), an open source platform for developing and running economic and behavioral experiments. The experiment is pre-registered at https://aspredicted.org/9gnf-s3j3.pdf. Data was analyzed using R (R Core Team, 2024) and RStudio (RStudio Team, 2024). All data, code, and materials, as well as video demos of the experiment are available at https://osf.io/fyehp/.

Design, Material and Procedure

We used a 2 (transparency vs. no transparency) $\times$ 2 (neutral framing vs. moral framing) fully between-subjects design. In all conditions, participants first received an introduction to the die-rolling game: They were informed that the player's task in this game is to report the outcome of die rolls as shown on screen over ten rounds, and that higher reported die rolls lead to higher payoffs (one cent per pip). These payoffs would be added to their study rewards as a bonus. They were also informed that rather than reporting the die roll outcomes themselves, an algorithm would report them on their behalf. Before proceeding to the main task and receiving more information about the algorithm, participants had to complete the following instruction check: "What determines the bonus payment in this task?", with the response options "actual die roll outcomes", "reported die roll outcomes", "chance", "average die roll outcomes". Participants were screened out if they didn't pick the correct response ("reported die roll outcomes") within two attempts.

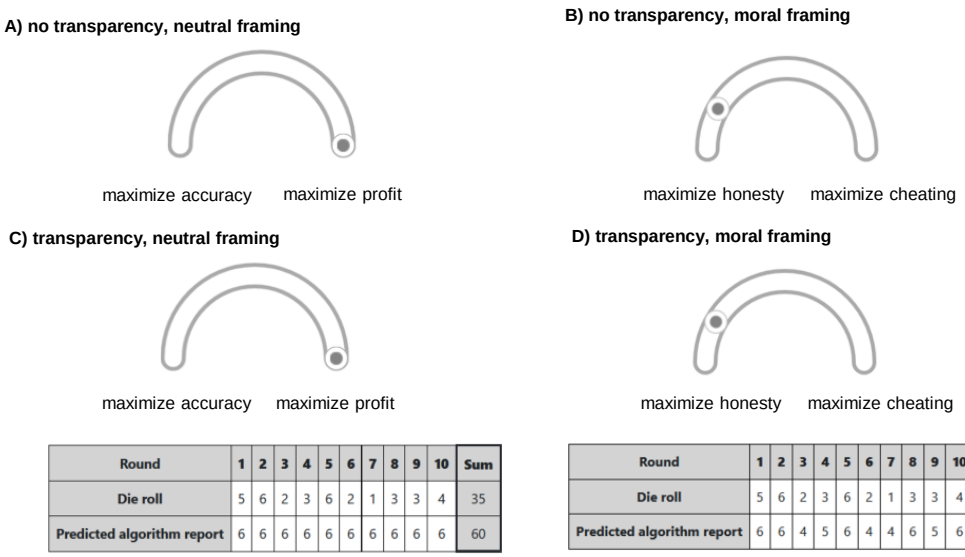
Main task: Delegation In all conditions, participants instructed an algorithm on how to report die roll outcomes on their behalf via a slider interface with an underlying 7-point response scale. In the "neutral framing" conditions (Fig. 1A and C), the slider endpoints were labeled with "maximize accuracy" (left) and "maximize profit" (right). In the "moral framing" conditions (Fig. 1B and D), the labels were "maximize honesty" (left) and "maximize cheating" (right). In the "no transparency" conditions (Fig. 1A and B), no further information about the algorithm was provided. In the "transparency" conditions (Fig. 1C and D), we provided a preview table that illustrated the algorithm's behavior for each setting of the slider and invited participants to explore as much as they wanted before committing on a final setting for the main task. The table contained a fixed sequence of ten hypothetical die rolls ("Die Roll"), and a second row that would display dynamically generated algorithm reports based on the currently chosen setting on the slider ("Predicted Algorithm Report"). The table also showed the sum of the actual and of the reported die rolls in a final column ("Sum"). Reported die rolls were generated using the same algorithm that would later be used in the main die-rolling game.

Algorithm The algorithm operated as follows: For each of the seven slider settings, a fixed target payoff sum for the ten rounds of the die-rolling game was defined—ranging from 35 (the expected value of fair, accurate reporting) to 60 (maximal over-reporting with all rolls as six). Intermediate target sums (39, 43, 48, 52, 56) were spaced between these extremes. Given ten candidate die roll reports, the algorithm checked if their sum matched the target sum for the selected setting. If not, it randomly altered one value and rechecked the sum, repeating this until the target was reached. This sequence was then reported. Since the example and main task die rolls were fixed to correspond to the fair expected value over 10 rolls (35), the algorithm did not alter any values for the first slider setting ("maximize accuracy"/"maximize honesty"), resulting in accurate reporting. This procedure aligns with Köbis et al. (2024).

Prediction task After picking a final setting for the algorithm and agreeing to start the die rolling game, participants in all conditions were additionally presented with a prediction task. We added this task to gauge participants' general payoff expectations across conditions, and to assess to what extent they had processed the information provided in the "transparency" conditions. Participants saw an image of the slider fixed to their previously chosen setting and a payoff table like the one used in the "transparency" conditions, but with an empty row below the ten hypothetical die rolls. Their task was to predict what they think the algorithm would report for each die roll, given the setting they had just selected for the algorithm. The displayed sequence of hypothetical die rolls was identical to the sequence displayed in the "transparency" conditions, and the sum of participants' predictions was automatically calculated and displayed in a final column. This task was incentivized for accuracy: We informed participants that if their predictions were in the top 10% in terms of accuracy, they would receive an additional bonus payment of £0.2.

Die rolls After the prediction task, participants proceeded to the main die rolling game. They observed ten videos of die rolls on separate pages, for each of which a reported die roll outcome by the algorithm would appear below the video frame after a short delay. The *actual* 10 die rolls were fixed (but displayed in random order). The *reported* die rolls were dynamically generated for each participant based on their chosen setting as explained above, and then sequentially displayed along with the die roll videos. After ten videos and algorithmic reports had been displayed, participants saw a summary table of the observed and the reported die rolls, and were informed about the size of their bonus payment. On a subsequent page, they were also informed about their achieved accuracy in the prediction task.

Lastly, participants completed an exit questionnaire comprised of a free text field asking them to explain how they completed the task ("In a few words, how did you decide which setting to pick for the algorithm?") as well as 17 lik-



Round	1	2	3	4	5	6	7	8	9	10	Sum
Die roll	5	6	2	3	6	2	1	3	3	4	35
Predicted algorithm report	6	6	6	6	6	6	6	6	6	6	60

Round	1	2	3	4	5	6	7	8	9	10	Sum
Die roll	5	6	2	3	6	2	1	3	3	4	35
Predicted algorithm report	6	6	4	5	6	4	4	6	5	6	52

Figure 1: Experimental materials per condition.

ert items asking about different aspects of their experience during the task (e.g., perceived permissibility of maximizing profit, perceived control, responsibility, and guilt, see OSF repository for the full text of all questions). The experiment ended with a debriefing and the assessment of demographic variables.

Participants

Sample size was determined by an a priori power simulation conducted with the *Superpower* package's (Lakens & Caldwell, 2021) web application (see https://shiny.ieis.tue.nl/anova_power/). We aimed for 800 participants, allowing us to detect main effects of at least $f = 0.02$ (small) for transparency or framing in a 2×2 ANOVA as well as a potential interaction of at least $f = 0.02$ (small) with more than 90% power each. 808 participants completed the survey on *prolific.com*. Inclusion criteria were: being a native English speaker, not having participated in previous studies using similar materials, and an approval rate of at least 90% on previous tasks on the platform. Participants received a compensation of £1.20 for an estimated 8 minutes of their time as a fixed payment, plus a bonus of at least £0.35 and up to £0.60 in the die rolling game. We excluded participants who failed an attention check (asking them to select "TikTok" as their main source of news from a list of platforms, regardless of what is actually true)[2], leaving us with a sample of 788 valid participants (M_{age} = 38.58, SD_{age} = 12.13 years, 46% women, 54% men, <1% non-binary or no answer).

[2]We preregistered a second attention check: "Please indicate your agreement with the statement below: Olive trees can communicate with each other using the internet", planning to exclude those who agreed with the statement. However, we later realized that this attention check is ambiguous and can be read as asking participants to agree with the statement as a means of passing the attention check. Thus, we did not exclude participants based on their answer to this question.

Results and Discussion

See Figure 2 for an overview of results for our main dependent measure, the setting people chose for the algorithm.

Preregistered analysis: Does transparency reduce cheating? Contrary to **Hypothesis 1**, providing transparency did not reduce requests for cheating in the die rolling game. With or without information about the algorithm's behavior given different settings, people on average selected a setting close to the scale midpoint (transparency: $M = 4.28$, $SD = 2.22$, no transparency: $M = 4.09$, $SD = 2.24$; $F_{1,784} = 1.10$, $p = 0.294$). The mode setting in both conditions however was 7 (chosen by 27% of participants in the "transparency" conditions, and by 24% in the "no transparency" conditions). That is, the largest chunk of people in these conditions requested maximal over-reporting. Another large cluster in both conditions set the algorithm up for fully accurate reporting (20% in the "transparency" conditions, 22% in the "no transparency" conditions). Finally, a third group of participants clustered around the midpoint of 4 (21% in the "transparency" conditions, 19% in the "no transparency" conditions). The average bonus payment achieved by participants was £0.49, SD = £0.09 with transparency and £0.48, SD = £0.09 without.

Framing, on the other hand, substantially reduced cheating in the die rolling game (**Hypothesis 2**). When the slider endpoints were labeled with "honesty" and "cheating", participants requested significantly less over-reporting ($M = 3.56$, $SD = 2.27$) than when they were labeled with "accuracy" and "profit" ($M = 4.80$, $SD = 2.02$, $F_{1,784} = 65.17$, $p < .001$, $f = 0.29$). The mode response was 7 (i.e., maximal over-reporting) with neutral framing (31% of responses), but it flipped to 1 (i.e., no over-reporting at all) with moral framing

(30% of responses). Nevertheless, there were also clusters around 7 (maximal over-reporting) in the "moral framing" conditions (20%), and around 1 (no over-reporting) in the "neutral framing" conditions (11%). Finally, 20% of participants opted for the midpoint in both conditions as well. The average bonus payment achieved by participants was £0.46, SD = £0.10, with moral framing and £0.51, SD = £0.08, with neutral framing. There was no significant interaction between transparency and framing ($F_{1,784} = 1.44$, $p = 0.23$).

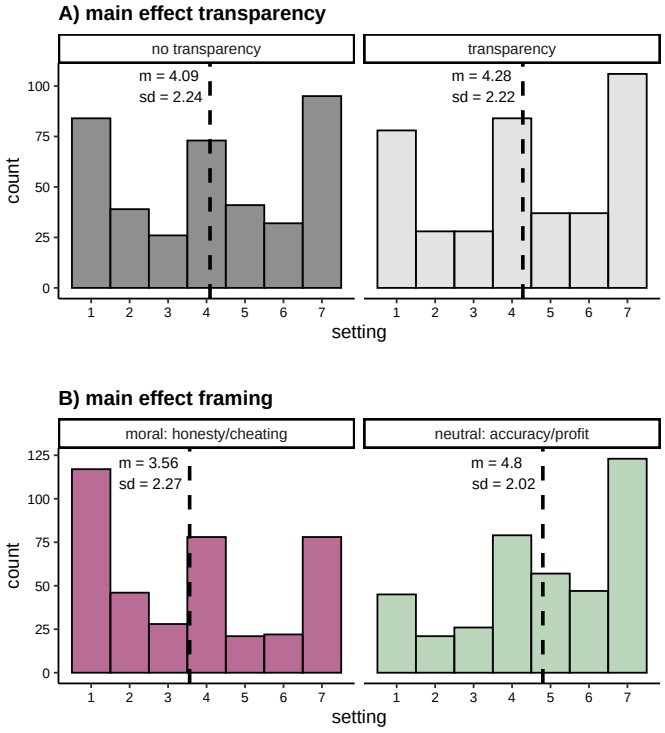


Figure 2: Main effects of mechanical (A) and framing (B) on the chosen setting for the algorithm.

Exploratory analysis 1: Do people seek and process transparency information? Given the surprising lack of effect for the transparency manipulation, we investigated whether participants assigned to these conditions a) sought and b) processed the information provided in the payoff tables. We tracked how often participants moved the slider before committing to a final setting, and observed that with transparency, people looked at on average $M = 7.52$ ($SD = 7.65$) settings before proceeding to the main task, but only at $M = 2.83$ ($SD = 2.69$) in the "no transparency" conditions ($F_{1,784} = 130.48$, $p < .001, f = 0.41$). Framing did not affect exploration behavior (moral framing: $M = 5.14$, $SD = 6$, neutral framing: $M = 5.25$, $SD = 6.43$, $F_{1,784} < .001$, $p = .99$), and there was no interaction ($F_{1,784} < .001$, $p = .98$). Thus, participants did seek transparency information when it was provided, generating predictions for the algorithm's behavior for all seven possible

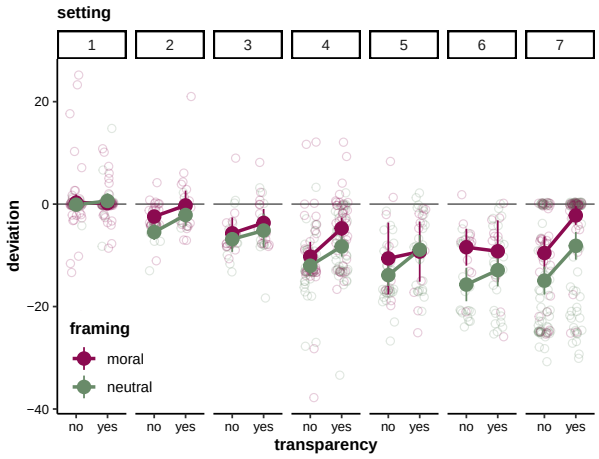


Figure 3: Mean deviation of participants' payoff predictions from actual payoffs per chosen setting for the algorithm, transparency and framing. Error bars indicate 95% confidence intervals.

settings on the slider on average.

Next, we explored whether people also *processed* the information they sought by analyzing their responses to the prediction task. Figure 3 summarizes the results of the prediction task by transparency and chosen setting (since predicting the algorithm's behavior is easier for those who chose an extreme setting, i.e., no over-reporting or full over-reporting). As Figure 3 shows and an exploratory 2 (transparency) $\times$ 2 (framing) $\times$ 7 (chosen setting) ANOVA corroborates, people generally underestimated profits, but to what extent they did so depended on their chosen setting ($F_{1,780} = 148.22$, $p < .001$, $f = .44$), transparency ($F_{1,780} = 46.01$, $p < .001$, $f = .24$), framing ($F_{1,780} = 41.42$, $p < .001$, $f = .23$), as well the two-way interactions transparency $\times$ setting ($F_{1,780} = 15.64$, $p < .001$, $f = .14$) and framing $\times$ setting ($F_{1,780} = 5.79$, $p = .016$, $f = .09$). Those who requested no over-reporting at all made accurate predictions about their payoffs, unaffected by any experimental manipulation (see "Setting = 1" in Figure 3). However, the higher the setting people had chosen for the algorithm, the more they underestimated profits. Accuracy was improved by the transparency manipulation, and more strongly for people who had selected a higher setting. Likewise, moral framing led to higher profit estimates, again more strongly for people who had selected a higher setting. Taken together, the results for this measure indicate that people did process the transparency information we provided, as it improved the accuracy of their profit predictions. In addition, however, predictions were also affected by the setting they had chosen for the algorithm and by the framing manipulation, with morally loaded language leading to higher profit expectations.

We obtained further evidence that people processed the transparency information by correlating the extent to which they explored the slider with the accuracy of their payoff pre-

dictions: Those who explored the slider more in the "transparency" conditions provided more accurate estimates in the prediction task (r = 0.21, p<.001), whereas no such correlation was observed in the "no transparency" conditions (r = -0.09, p = .066).

Exploratory analysis 2: How did people experience the task? In their free-text explanations, participants emphasized their motives: a desire for honesty (e.g., "honesty is worth more than money to me"), a preference to maximize payoff (e.g., "I wanted to receive the maximum bonus"), or an attempt to strike a balance (e.g., "to maximise the bonus without straying too far from accuracy"). The frequency of mention aligns with the behavioral results: A desire for honesty was more often mentioned with moral framing than neutral framing ($\chi^2_{df=1}$ = 16.7, p<.001), while maximizing profit was more often mentioned with neutral framing ($\chi^2_{df=1}$ = 6.23, p = .013). In line with the null effect for transparency, we only occasionally observed that people cited studying the payoff table as the basis for their behavior (e.g., "I analyzed the table").

On the exit questionnaire, participants indicated that it seemed more morally permissible to them to pick a high setting for the algorithm with neutral framing (M = 3.89, SD = 1.10) than with moral framing (M = 3.26, SD = 1.37, $F_{1,785}$ = 49.92, p <.001, f = 0.25), and they also indicated that they felt more honest in the "neutral framing" condition (M = 4.32, SD = 1.01) than with moral framing (M = 3.85, SD = 1.35, $F_{1,785}$ = 30.87, p <.001, f = 0.20). Neither of these measures was affected by the transparency manipulation. However, when asked whether they felt guilty for the reported die roll outcomes, for how the algorithm reported the die roll outcomes on their behalf, or for the size of their bonus payment, ratings were low across the board, and unaffected by any manipulation. Thus, the exit questions largely corroborated the pattern of effects observed for our main dependent measure. The full results for all exit questions are provided in the OSF repository.

General Discussion

We investigated the effects of two interventions as means of curbing cheating in delegation to AI: transparency, operationalized as communicating *how much* unearned gain an algorithmic delegate produces for the user and *how* it achieves this, and framing, operationalized as making the ethical significance of user behavior explicit using natural language cues. We found that that transparency is not sufficient to curb cheating. Participants were just as likely to request over-reporting in the die-rolling game whether or not they had full visibility into the algorithm's mechanics and the consequences of their choices. Despite the null effect on behavior, process measures revealed that participants actively engaged with transparency information. They explored and processed the details provided, which suggests that the lack of behavioral change was not due to disinterest or misunderstanding. Instead, the manipulation appears insufficient to shift participants' moral interpretation of the task or to motivate ethical behavior. In contrast, making the ethical dimension of the task transparent significantly influenced behavior. Explicitly labeling dishonest actions as such reduced cheating, likely by heightening participants' moral discomfort, as suggested by exit questionnaire data. This aligns with calls for a broader understanding of transparency, encompassing both ethical and technical dimensions (Ananny & Crawford, 2018; Kolkman, 2022; Felzmann et al., 2020; Andrada et al., 2023; Felzmann et al., 2019). Clear, interpretive aids—such as natural language descriptions—may be essential to help users understand the moral implications of algorithmic behavior.

A limitation of the current design is the coupling of participants' predictions about the algorithm's behavior with their own choice of algorithmic settings. This may have led participants to provide biased estimates, potentially as a form of self-justification or retaining plausible deniability of an intention to cheat. Future research should separate estimation tasks from decision-making to disentangle these effects and better understand how people process transparency information.

In sum, we show that informing users about an algorithm's mechanics and outcomes does not necessarily deter unethical behavior. Instead, elements like morally loaded natural language cues appear to play a more significant role. Future studies should explore in more detail how these dimensions affect user understanding, ideally across a wider range of tasks and incorporating more advanced AI systems, such as Large Language Models.

References

Akhtar, M. A. K., Kumar, M., & Nayyar, A. (2024). Transparency and accountability in explainable AI: Best practices. In *Towards ethical and socially responsible explainable ai: Challenges and opportunities* (pp. 127–164). Springer.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.

Andrada, G., Clowes, R. W., & Smart, P. R. (2023). Varieties of transparency: Exploring agency within AI systems. *AI Society*, *38*(4), 1321–1331.

Bandura, A. (2002). Selective moral disengagement in the exercise of moral agency. *Journal of Moral Education*, *31*(2), 101–119.

Bartling, B., & Fischbacher, U. (2012). Shifting the blame: On delegation and responsibility. *The Review of Economic Studies*, *79*(1), 67–87.

Candrian, C., & Scherer, A. (2022). Rise of the machines: Delegating decisions to autonomous AI. *Computers in Human Behavior*, *134*, 107308.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97. doi: 10.1016/j.jbef.2015.12.001

Cheong, B. C. (2024). Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, *6*, 1421273.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition*, *108*(2), 353–380.

Engelmann, N., & Waldmann, M. R. (2022). How causal structure, causal strength, and foreseeability affect moral judgments. *Cognition*, *226*, 105167.

Feier, T., Gogoll, J., & Uhl, M. (2021). Hiding behind machines: When blame is shifted to artificial agents. *arXiv preprint arXiv:2101.11465*.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333–3361.

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, *6*(1), 2053951719860542.

Fischbacher, U., & Föllmi-Heusi, F. (2013). Lies in disguise—an experimental study on cheating. *Journal of the European Economic Association*, *11*(3), 525–547.

Hamman, J. R., Loewenstein, G., & Weber, R. A. (2010). Self-interest through delegation: An additional rationale for the principal-agent relationship. *American Economic Review*, *100*(4), 1826–1846.

Hill, A. (2015). Does delegation undermine accountability? Experimental evidence on the relationship between blame shifting and control. *Journal of Empirical Legal Studies*, *12*(2), 311–339.

Hohenstein, J., & Jung, M. (2020). AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior*, *106*, 106190.

Illia, L., Colleoni, E., & Zyglidopoulos, S. (2023). Ethical implications of text generation in the age of artificial intelligence. *Business Ethics, the Environment & Responsibility*, *32*(1), 201–210.

International Monetary Fund. (2024). *Advances in artificial intelligence: Implications for capital market activities*. Washington, DC: International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/GFSR

Kolkman, D. (2022). The (in) credibility of algorithmic models to non-experts. *Information, Communication & Society*, *25*(1), 93–109.

Köbis, N., Rahwan, Z., Bersch, C., Ajaj, T., Bonnefon, J.-F., & Rahwan, I. (2024). *Experimental evidence that delegating to intelligent machines can increase dishonest behaviour.* (Preprint) doi: 10.31219/osf.io/dnjgz

Lagnado, D. A., & Gerstenberg, T. (2017). Causation in legal and moral reasoning. In *Oxford handbook of causal reasoning* (pp. 565–602). Oxford University Press.

Lakens, D., & Caldwell, A. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1), 251524592095150. doi: 10.1177/2515245920951503

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy & Technology*, *31*(4), 611–627.

MacKay, A., & Weinstein, J. (2022). Dynamic pricing algorithms, consumer harm, and regulatory response. *Washington University Law Review*, *100*(1), 23–68.

Moore, C. (2015). Moral disengagement. *Current Opinion in Psychology*, *6*, 199–204.

Newman, A., Le, H., North-Samardzic, A., & Cohen, M. (2020). Moral disengagement at work: A review and research agenda. *Journal of Business Ethics*, *167*, 535–570.

Oexl, R., & Grossman, Z. J. (2013). Shifting the blame to a powerless intermediary. *Experimental Economics*, *16*, 306–312.

Paharia, N., Kassam, K. S., Greene, J. D., & Bazerman, M. H. (2009). Dirty work, clean hands: The moral psychology of indirect agency. *Organizational Behavior and Human Decision Processes*, *109*(2), 134–141.

Parlangeli, O., Curro', F., Palmitesta, P., & Guidi, S. (2024). Moral judgements of errors by AI systems and humans in civil and criminal law. *Behaviour & Information Technology*, *43*(9), 1718–1728.

R Core Team. (2024). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Royzman, E. B., & Baron, J. (2002). The preference for indirect harm. *Social Justice Research*, *15*, 165–184.

RStudio Team. (2024). RStudio: Integrated Development Environment for R [Computer software manual]. Boston, MA. Retrieved from https://posit.co/

Shank, D. B., DeSanti, A., & Maninger, T. (2019). When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions. *Information, Communication & Society*, *22*(5), 648–663.

Waldmann, M. R., Wiegmann, A., & Nagel, J. (2017). Causal models mediate moral inferences. In *Moral inferences* (pp. 45–63). Psychology Press.

Weiss, A., & Forstmann, M. (2024). Religiosity predicts the delegation of decisions between moral and self-serving immoral outcomes. *Journal of Experimental Social Psychology*, *113*, 104605.

Ziano, I., Wang, Y. J., Sany, S. S., Ngai, L. H., Lau, Y. K., Bhattal, I. K., ... others (2021). Perceived morality of direct versus indirect harm: Replications of the preference for indirect harm effect. *Meta-Psychology*, *5*.