

Confidence Amplification in Human–AI Moral Decision-Making

Proceedings of the Human Factors and
Ergonomics Society Annual Meeting
1–4

© The Author(s) 2026



Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/10711813261493608
journals.sagepub.com/home/pro



Mengyao Li¹, Trevor Patten¹, Nishthaa Lekhi², and Areen Alsaid²

Abstract

As artificial intelligence (AI) systems increasingly engage users in value-laden discussions, a key concern is whether they shape not only what people decide but how confident they feel. We examined confidence amplification, shifts in certainty without decision reversals, in human–AI moral decision-making. In a 2 (Time: pre vs. post) × 3 (AI moral framing: utilitarian, deontological, balanced) mixed design, 120 participants engaged a 10-round conversation with an AI assistant discussion an autonomous-vehicle dilemma. Decision reversals were uncommon and unaffected by framing. However, confidence increased reliably from pre- to post-discussion, and a confidence-weighted opinion-strength measure shifted toward the AI's framing. Echo-chamber dynamics emerged when framing aligned with participants' initial stance: aligned participants gained confidence whereas misaligned participants did not, with parallel increases in linguistic certainty. These findings suggest that amplification of confidence, rather than directional opinion change, may be a central mechanism through which AI influences moral judgment.

Keywords

human–AI interaction, moral dilemma, decision-making, AI influence, large language models

Introduction

People don't just ask artificial intelligence (AI) systems for purely factual information: in real-world deployments, everyday prompts routinely elicit value-laden responses (Huang et al., 2025). For instance, an employee asks whether to report a colleague's misconduct. Should the AI emphasize organizational or societal harm, or workplace harmony? When used in domains that include ethically consequential judgment, a key concern is that such systems may influence not only what decisions people make, but also how certain they feel about those decisions, and confidence can affect downstream behaviors, such as willingness to act, persistence under counterarguments, and propagation to others. Most work on moral influence is framed as categorical opinion change (e.g., switching between utilitarian and deontological choices). Here we focus on a subtler but consequential pathway: *confidence amplification* without decision reversals. In this paper, we conducted a study in which participants discussed an autonomous-vehicle (AV) moral dilemma with an AI assistant and investigated the effects of AI bias on people's opinion change and their confidence.

Background

People often apply social heuristics to interactive technologies, responding to them as social actors rather than neutral

tools (Moon & Nass, 1996). In judgment settings, people can rely heavily on algorithmic advice (Logg et al., 2019). Prior studies have shown adding explanations to the AI decisions does not appear to reduce the overreliance: people still on average make poorer final decisions than they would have without AI's assistance (Bansal et al., 2021; Bućinca et al., 2020, 2021). Thus, understanding LLM's influences on human moral-decision-making is critical yet underexplored. Moral dilemmas provide a structured way to study how framing shapes judgments: trolley-type tasks contrast utilitarian choices (maximizing aggregate welfare) with deontological choices (upholding duties and prohibitions), and AV variants show both cross-cultural regularities and divergence (Awad et al., 2018). Importantly, influence need not appear as decision switching: attitude certainty can change even when choice direction remains stable, with downstream consequences for persistence and behavior (Clarkson et al., 2008; Petrocelli et al., 2007). These dynamics suggest an echo-chamber pathway in which aligned AI framings amplify

¹Georgia Institute of Technology, Atlanta, USA

²University of Michigan, Dearborn, USA

Corresponding Author:

Mengyao Li, Georgia Institute of Technology, Atlanta, GA 30332, USA.
Email: mengyao.li@gatech.edu

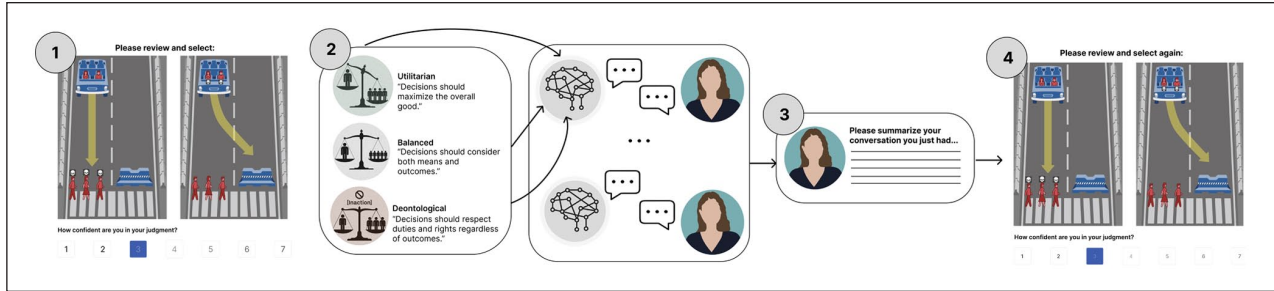


Figure 1. Experiment flow.

certainly, whereas misaligned framings may produce weaker gains or resistance. We therefore ask:

1. How do AI bias conditions (utilitarian, deontological, balanced) affect opinion change and confidence in moral dilemmas?
2. Does human–AI moral alignment amplify opinion reinforcement and confidence?

Method

We adopted a 2 (Time: pre vs. post; within-subjects) $\times$ 3 (AI moral framing: utilitarian, deontological, balanced; between-subjects) mixed design. We recruited $N=121$ U.S.-based English-speaking participants via Prolific. One participant failing an attention check was excluded, yielding $N=120$ evenly distributed across conditions, with a mean age of 38.6 ($SD=10.4$) and 40.8% identified as female. After providing informed consent and completing a commitment agreement to respond truthfully, participants provided demographic information and proceeded to the experimental task shown in Figure 1. Participants completed an AV moral dilemma adapted from Moral Machine (Awad et al., 2018), choosing between (i) swerving into a barrier (kill fewer pedestrians as utilitarian) versus (ii) continuing forward (without active harmful actions as deontological), and then engaged in a structured 10-round dialogue with an AI assistant prompted to adopt utilitarian (persuasively outcome-maximizing), deontological (persuasively duty/rights-based), or balanced (arguments for both sides) framing. For *LLM moral framing*, all prompts were pretested for perceived bias direction, emotional tone, and rhetorical strength. Conversation started with LLM reinstating the situation and asking for users' initial thoughts. After users presented their perspectives, the LLM responded using pre-tested bias statements. For example, "While it's understandable to consider who 'chose' to be in the car, a deontological perspective emphasizes that intentionally swerving into the barrier would be to directly cause harm to the passengers." Participants provided at least five responses, each at least 100 characters, leading to 10-round

conversations from users and LLM. At pre- and post-conversation timepoints, participants reported (a) a binary decision (utilitarian vs. deontological) and (b) confidence on a 7-point Likert scale to capture directional change without requiring reversals, we computed an *opinion strength* score $s \in [-7, 7]$ by assigning $+c$ to utilitarian decisions and $-c$ to deontological decisions, where c is confidence, and defined opinion shift as $\Delta s = s_{\text{post}} - s_{\text{pre}}$.

Results

First, we assessed decision reversals using a chi-square test and analyzed confidence and opinion strength using linear mixed-effects models with fixed effects of *Time* (pre vs. post), *Bias* condition, and their interaction, with participant-level random intercepts and Time random slopes. Overall, participants entered the study with a strong utilitarian preference (109/120; 90.8%). After the discussion, switching was uncommon (14/120; 11.7%) and did not differ by framing, $\chi^2(2)=1.13$, $p>.05$.

Despite stable choices, as shown in Figure 2, confidence increased reliably from pre ($M=5.65$, $SE=0.25$) to post ($M=6.23$, $SE=0.25$). The *Time* effect was significant, $\beta=0.58$, $z=2.86$, $p=.004$, $d=0.40$. A confidence-weighted opinion-strength measure captured more subtle influence: the *Time* $\times$ deontological-framing interaction was significant, $\beta=-2.23$, $z=-2.44$, $p=.015$, $d=-0.38$, and utilitarian framing produced a within-condition increase, $t(39)=3.03$, $p_{\text{adj}}=.004$, $d=0.48$.

Finally, evidence for echo-chamber dynamics emerged when framing aligned with participants' initial stance. Excluding the balanced condition ($n=80$), the aligned group significantly increase confidence from pre- to post-discussion, $\beta=0.92$, $p_{\text{adj}}<.001$, whereas misaligned group didn't, $p_{\text{adj}}>.05$. Dialogue language showed a parallel pattern: users expressed higher linguistic certainty than the AI overall, $\beta=0.422$, $z=6.80$, $p<.001$, and certainty was higher in biased framings (vs. balanced): deontological, $\beta=0.263$, $p_{\text{adj}}<.001$; utilitarian, $\beta=0.193$, $p_{\text{adj}}=.001$.

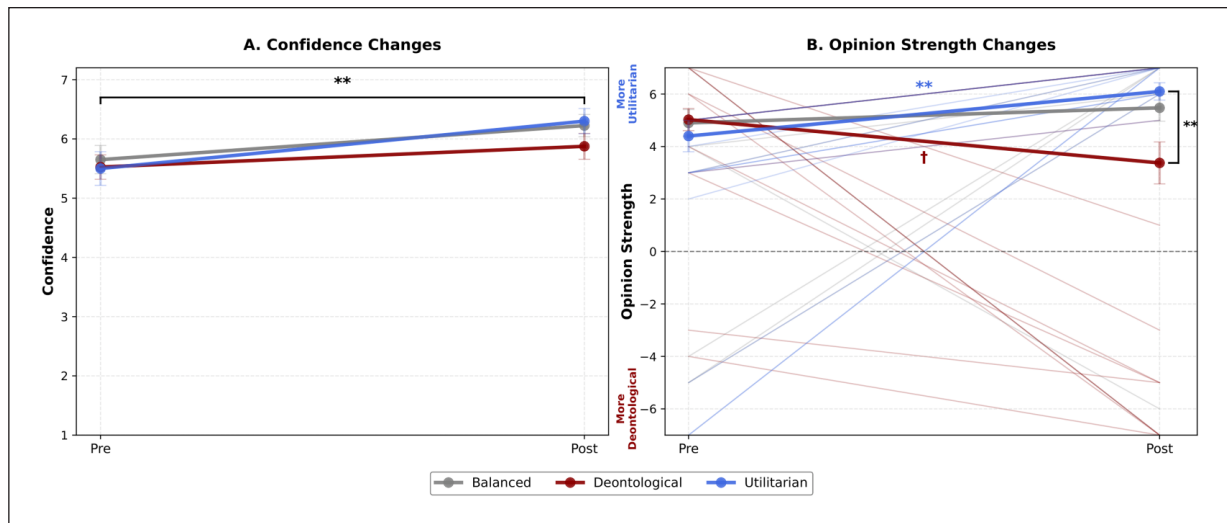


Figure 2. Confidence and opinion strength changes by AI moral framing: (A) shows confidence increased For Peer Review uniformly across all conditions from pre- to post-discussion and (B) shows opinion strength changes (-7 = deontological, $+7$ = utilitarian).

Note. Individual trajectories shown for participants with $|\Delta| \geq 2$.

** $p < .01$. † $p < .05$.

Conclusion

Across moral framings, AI-assisted dialogue produced reliable *confidence amplification* with relatively few categorical reversals. A confidence-weighted opinion strength measure shifted toward the AI's framing. Moreover, alignment between participants' initial stance and the AI's framing further amplified confidence, with parallel increases in linguistic certainty during dialogue.

This pattern matters because confidence functions as a social signal: high-certainty judgments are more likely to be acted upon, repeated, and defended, even when the underlying choice is unchanged. These findings extend evidence that AI-generated arguments are persuasive, particularly when they are personalized to characteristics of the recipient (Matz et al., 2024; Salvi et al., 2025), and converge with work on human–AI feedback loops that reshape collective judgments over time (Burton et al., 2024; Glickman & Sharot, 2024). Our results suggest that amplification of confidence, rather than directional change alone, may be a central mechanism through which AI systems exert influence.

Implications and Future Studies

Confidence functions as a socially consequential signal with broader societal implications. High-certainty judgments are more likely to be publicly expressed, selectively shared, and resistant to counterevidence. In a human–LLM hybrid information ecosystem characterized by repeated daily exposure, aligned interactions may incrementally reinforce certainty, facilitating echo-chamber dynamics and accelerating opinion cascades that deserve further investigation and proper risk mitigation (Pulford et al., 2018; Tsvetkova et al., 2024).

This study is limited by its use of artificial ethical dilemma scenarios within a U.S. sample in a single-session interaction. Because participants entered with a predominantly utilitarian stance, the echo-chamber effect was estimated largely from utilitarian-aligned participants, so future work should recruit a more balanced range of initial moral preferences to test how broadly this pattern generalizes. Future work should also test whether confidence amplification persists over repeated exposure and can be mitigated with interventions.

Acknowledgments

We would like to thank Shaurya Kumar for developing the study platform for this project.

ORCID iDs

Mengyao Li  <https://orcid.org/0000-0002-0819-4693>

Areen Alsaid  <https://orcid.org/0000-0003-2852-9750>

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

- Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, 563(7729), 59–64. <https://doi.org/10.1038/s41586-018-0637-6>

- 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* [Conference session]. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–16. ACM Press.
- Buçinca, Z., Lin, P., Gajos, K. Z., & Glassman, E. L. (2020). *Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems* [Conference session]. Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 454–464. ACM Press.
- 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. <https://doi.org/10.1145/3449287>
- Burton, J. W., Lopez-Lopez, E., Hechtlinger, S., Rahwan, Z., Aeschbach, S., Bakker, M. A., Becker, J. A., Berditchevskaia, A., Berger, J., Brinkmann, L., Flek, L., Herzog, S. M., Huang, S., Kapoor, S., Narayanan, A., Nussberger, A. M., Yasseri, T., Nickl, P., Almaatouq, A., ... & Hertwig, R. (2024). How large language models can reshape collective intelligence. *Nature Human Behaviour*, 8(9), 1643–1655. <https://doi.org/10.1038/s41562-024-01959-9>
- Clarkson, J. J., Tormala, Z. L., & Rucker, D. D. (2008). A new look at the consequences of attitude certainty: The amplification hypothesis. *Journal of Personality and Social Psychology*, 95(4), 810. <https://doi.org/10.1037/a0013192>
- Glickman, M., Sharot, T. How human–AI feedback loops alter human perceptual, emotional and social judgements. (2025). *Nature Human Behaviour*, 9(2), 345–359. <https://doi.org/10.1038/s41562-024-02077-2>
- Huang, S., Durmus, E., McCain, M., Handa, K., Tamkin, A., Hong, J., Stern, M., Somani, A., Zhang, X., & Ganguli, D. (2025). Values in the wild: Discovering and analyzing values in real-world language model interactions. *arXiv Preprint arXiv:2504.15236*.
- 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. <https://doi.org/10.1016/j.obhdp.2018.12.005>
- Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M., & Cerf, M. (2024). The potential of generative AI for personalized persuasion at scale. *Scientific Reports*, 14(1), 4692. <https://doi.org/10.1038/s41598-024-53755-0>
- Moon, Y., & Nass, C. (1996). How “real” are computer personalities? Psychological responses to personality types in human-computer interaction. *Communication Research*, 23(6), 651–674. <https://doi.org/10.1177/009365096023006002>
- Petrocelli, J. V., Tormala, Z. L., & Rucker, D. D. (2007). Unpacking attitude certainty: Attitude clarity and attitude correctness. *Journal of Personality and Social Psychology*, 92(1), 30. <https://doi.org/10.1037/0022-3514.92.1.30>
- Pulford, B. D., Colman, A. M., Buabang, E. K., & Krockow, E. M. (2018). The persuasive power of knowledge: Testing the confidence heuristic. *Journal of Experimental Psychology: General*, 147(10), 1431. <https://doi.org/10.1037/xge0000471>
- Salvi, F., Horta Ribeiro, M., Gallotti, R., & West, R. (2025). On the conversational persuasiveness of GPT-4. *Nature Human Behaviour*, 9(8), 1645–1653. <https://doi.org/10.1038/s41562-025-02194-6>
- Tsvetkova, M., Yasseri, T., Pescetelli, N., & Werner, T. (2024). A new sociology of humans and machines. *Nature Human Behaviour*, 8(10), 1864–1876. <https://doi.org/10.1038/s41562-024-02001-8>