Computers in Human Behavior Reports 24 (2026) 101473

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Shifting confidence, not choices: Moral bias and social influence in human-LLM conversations

Mengyao Li [a,*], Trevor Patten [a], Nishthaa Lekhi [b], Areen Alsaid [b]

[a] School of Psychological and Brain Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA
[b] Industrial and Manufacturing System Engineering, University of Michigan–Dearborn, Dearborn, MI, 53706, USA

ARTICLE INFO

Keywords:
Large language models
Moral decision-making
Human–AI interaction
Social influence
Information cascade

ABSTRACT

Large language models (LLMs) are increasingly embedded in everyday decision-making, yet their influence on human moral judgment remains poorly understood. Prior research has largely tracked binary opinion reversal as the primary outcome, overlooking a more consequential pathway: modulation of the certainty with which moral positions are held, which can shape downstream behaviors and susceptibility to social influence. To test this, we conducted two experiments (N = 120 each) in which online samples of US adults (M_{age} = 38.6, SD = 10.4, 40.8% female) conversed with a morally framed LLM (utilitarian, deontological, or balanced) about autonomous vehicle moral dilemmas. Experiment 2 extended this to a two-stage social cascade in which downstream participants received either LLM- or human-generated summaries of prior human–LLM conversations. Results showed that binary moral decisions were highly stable. Confidence increased across all three bias conditions in both experiments, and confidence-weighted opinion strength shifted toward the LLM's framing. In the predominantly utilitarian sample, alignment was associated with confidence gains, whereas misalignment shifted opinion direction without reducing confidence. Experiment 2 used highly aligned summaries to model a high-risk cascade. Surprisingly, moral-bias effect sizes were descriptively smaller in Experiment 2 than in Experiment 1, although the between-experiment interaction was not significant. Deontological framing was descriptively more resilient, and AI-generated summaries produced a stronger deontological shift than human-generated summaries. In these experiments, LLM dialogue influenced confidence even when choices remained stable; whether this pattern generalizes to other models, moral dilemmas, and populations remains to be tested.

1. Introduction

Large language models (LLMs) are increasingly embedded in everyday decision-making, not merely as information retrieval tools but as interlocutors that generate morally laden arguments, frame trade-offs, and engage users in extended deliberation (Huang et al., 2025). Moral issues inevitably appear in conversations with chatbots due to their prevalence in everyday scenarios people want advice on (e.g., "Should I tell my manager that a teammate has been quietly overstating billable hours on a client project?"). Critically, LLMs are not value-neutral: they exhibit emergent value systems shaped by training data, optimization objectives, and alignment procedures (Mazeika et al., 2025), which can remain internally coherent even under instructions to remain neutral. For instance, when a clinician asks an AI whether to allocate a scarce treatment to one patient over another, the system does not simply retrieve facts. It frames what considerations matter, how much weight to assign them, and how confident a person should feel in acting on the conclusion. In these conversations, people apply social heuristics to LLMs, treating them as social actors rather than neutral tools (Nass & Moon, 2000). These properties raise a fundamental question that existing research has yet to address directly: how do LLMs influence what people decide and how certain they feel about the decisions they already hold in moral dilemmas?

Most empirical work on AI moral influence operationalizes persuasion as categorical opinion reversal: whether participants switch positions after exposure to AI-generated arguments (Fabre et al., 2024). This framing, while tractable, risks overlooking a more socially consequential pathway. Moral attitudes are among the most resistant to direct persuasion (Luttrell et al., 2016; Skitka et al., 2021) and brief AI interactions rarely overturn them. But the same interactions may systematically amplify the confidence with which initial positions are held. Because attitude certainty predicts behavioral consistency, persistence

* Corresponding author.
E-mail addresses: mengyao.li@gatech.edu (M. Li), tpatten7@gatech.edu (T. Patten), nlekhi@umich.edu (N. Lekhi), alsaid@umich.edu (A. Alsaid).

https://doi.org/10.1016/j.chbr.2026.101473
Received 4 May 2026; Received in revised form 28 October 2026; Accepted 18 November 2026
Available online 23 November 2026
2451-9588/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

under counterargument, and willingness to advocate publicly (Clarkson et al., 2008), confidence amplification, even without decision change, constitutes a meaningful and understudied form of influence. To capture this, the present research introduces a confidence-weighted opinion strength metric that integrates decision direction and certainty into a single scalar, enabling the detection of influence that binary coding would miss.

A second gap concerns the social propagation of LLM-induced confidence. LLMs do not operate in isolation: their outputs are summarized, shared, and recirculated in digital environments. When downstream users encounter AI-curated summaries of prior human–LLM conversations, they are exposed not only to moral content, but to the confidence with which that content was framed. Whether LLM-amplified moral confidence propagates through such social cascades and whether the source of summarization (human vs. AI) modulates this transmission remain unexplored. This question is consequential: if AI-generated summaries preserve moral framing more faithfully than human-generated ones, AI-mediated information chains may function as confidence-hardening pipelines that progressively narrow the space for genuine moral deliberation.

To address these gaps, we report two experiments using autonomous vehicle moral dilemmas adapted from the Moral Machine framework (Awad et al., 2018). **Experiment 1** exposed participants to a structured conversation with a morally biased LLM, framed as utilitarian, deontological, or balanced, and measured pre-to-post changes in binary moral decisions, confidence and opinion strength. **Experiment 2** extended this to a two-stage social cascade in which participants first read a summary of a prior human–LLM conversation, then engaged in their own LLM discussion. Summaries were authored either by a human participant or by the LLM itself, enabling examination of how LLM-introduced bias propagates across cascade positions and whether the summarizer's identity moderates downstream bias transmission. Together, these experiments examine how an LLM's moral framing influences human–LLM conversations and inform implications for AI transparency and the governance of human–AI decision systems.

2. Background

2.1. LLMs as active moral influencers in dialogue

Trolley problems offer a structured paradigm for studying moral judgment and decision-making by contrasting utilitarian judgments (maximizing aggregate welfare) with deontological judgments (upholding duties regardless of consequences) (Thomson, 1985). The Moral Machine framework extended this paradigm to autonomous-vehicle dilemmas at global scale, showing that preferences over whom a vehicle should spare are systematic yet culturally variable (Awad et al., 2018), which establishes that such dilemmas elicit stable moral positions that an interlocutor could plausibly shift. LLMs are well positioned to exert such influence because they carry emergent value systems, built-in value leanings that show up even when their answers sound neutral (Mazeika et al., 2025), which positions them as active moral influencers rather than passive tools. Past research has shown that AI moral advice reliably shifts judgments. Fabre et al. (2024) found that utilitarian or deontological arguments from an AI advisor significantly influenced participants' decisions on sacrificial dilemmas, and Krügel et al. (2023, 2026) showed that a single ChatGPT recommendation moved users' moral judgments, even when the advice was disclosed as machine-generated and whether or not it carried a justification. Yet, these demonstrations share two features that limit what they reveal about naturalistic influence. First, each delivers the model's perspective as a static, one-shot input presented before the participant commits, rather than engaging the model as an active interlocutor that elaborates, responds, and negotiates across turns, the mode in which people increasingly consult these systems. Second, each index influence as a change in the decision itself, leaving unexamined how strongly a position is held. Understanding influence in dialogue therefore requires both an interactive paradigm and a specification of what should count as influence when moral positions rarely reverse.

2.2. Confidence as the pathway of influence

Moral attitudes are among the most resistant to direct persuasion: attitudes held with moral conviction are experienced as universally and objectively true and are comparatively immune to authority and peer influence (Skitka et al., 2021). Resistance to reversal, however, does not mean the attitude is untouched. Attitude certainty describes how firmly a position is held, indexed by properties such as confidence, whereas moral conviction describes the perceived basis of the position, the sense that it follows from a fundamental distinction between right and wrong (Skitka et al., 2005). Luttrell et al. (2016) further support treating perceived moral basis as a separate indicator of attitude strength. Influence can therefore register on strength while direction stays fixed, and certainty is the most tractable index of that shift, because it governs how strongly a position shapes judgment and behavior, amplifying the attitude's dominant effect rather than uniformly hardening it (Clarkson et al., 2008). In addition, LLMs introduce a further complication: models trained with reinforcement learning from human feedback (RLHF) exhibit sycophancy, systematically affirming rather than challenging user views, which can further raise confidence and alter people's behavioral inclinations (Cheng et al., 2026). Together, these findings reframe what should count as influence in human-LLM moral dialogue. If LLM influence operates through confidence rather than choice, binary coding of decision reversals will miss the primary signal, because people may hold the same decision while changing how strongly it is held, a distinction that requires separating attitude direction from attitude certainty. We therefore operationalize influence as a confidence-weighted opinion strength metric that integrates decision direction and certainty.

When chatting with LLMs, influence can operate through subtle reframing and incremental shifts across conversational turns rather than overt persuasion in behavioral and self-report measures. Thus, analyzing linguistic markers of certainty in human-LLM dialogue can complement subjective measures by revealing process: how participants frame their reasoning, how certainty is expressed, and how language shifts across turns in a dialogue. Dictionary-based (Pennebaker et al., 2015) and epistemic-stance analyses of hedges and certainty markers (Hyland, 1998; Lakoff, 1975) provide a process-level window on how conviction shifts across conversational turns, complementing self-report. How certainty language trajectories evolve across human–LLM moral conversations and what they reveal about the conversational mechanisms of confidence amplification has not previously been examined.

2.3. From dyad to cascade: confidence propagation and summarizer roles

LLM influence does not remain contained to a single conversation. LLM outputs are summarized, shared, and recirculated, creating diffusion chains in which each person's rendering of a message becomes the next person's input (Moussaïd et al., 2015). Summarization carries influence across each link by preserving what is salient and signaling which ideas deserve emphasis (Lorch et al., 2001), but each link is a selective act rather than a neutral relay: people summarize and forward content aligned with their prior views, forming echo chambers (Cinelli et al., 2021; Hart et al., 2009), and moralized content is especially prone to diffuse (Brady et al., 2017). What propagates through such chains is therefore not only moral content, but the certainty attached to it. Certainty does not travel as a uniform hardening agent: under the amplification hypothesis, certainty amplifies an attitude's dominant effect rather than invariably strengthening it, hardening univalent attitudes but weakening ambivalent ones (Clarkson et al., 2008). Inherited certainty should thus intensify whatever leaning a downstream reader already holds, so cascade positions occupied by more ambivalent

readers should attenuate rather than compound.

The identity of the summarizer, human or LLM, can influence summarization and confidence propagation. On the reception side, authorship acts as an agency cue: the machine heuristic leads readers to treat machine-authored content as objective and free of ideological bias (Metzger & Flanagin, 2007; Sundar & Kim, 2019, pp. 1–9), although explicit AI labels can instead depress perceived accuracy and willingness to share identical content (Altay & Gilardi, 2024). On the production side, human summarizers are subject to the stance-consistent selection described above, whereas an LLM has no prior moral attitude to protect and may preserve source framing more faithfully. Whether summarizer identity strengthens or dampens the downstream transmission remains underexplored.

Beyond who does the summarizing, the presence of an LLM in the loop is itself amplifying. Glickman and Sharot (2025) found that human–AI feedback loops amplified human bias more than interacting with another person did. Lu and colleagues (2022) showed that AI credibility cues shift the spread of misinformation more when social influence is present than when it is absent. Yet, this evidence comes mostly from misinformation, where factual benchmarks let claims be verified as true or false. Moral dilemmas lack such a corrective anchor, so AI-amplified confidence may constitute unwarranted affirmation with nothing to constrain it, making moral cascades potentially more consequential. How moral confidence itself propagates through such cascades, however, remains unexamined.

2.4. The present study

Taken together, as LLMs increasingly participate in ethically consequential dialogue, it is imperative to examine how LLM moral bias shapes decision-making in direct interaction and across a social cascade, and whether the source of summarization moderates downstream transmission. Experiment 1 tests whether and how LLM moral bias affects humans' moral decisions and confidence in direct conversation and whether alignment between a participant's prior view and the model's framing amplifies confidence. Experiment 1 asks:

1. **RQ1**: How do different LLM bias conditions (utilitarian, deontological, balanced) influence opinion change and confidence in moral decision-making in social dilemmas?
2. **RQ2**: Does alignment between a person's moral view and an LLM's bias (i.e., echo chamber) amplify the person's opinion strength and confidence?

Experiment 2 then examines how LLM-introduced moral bias propagates across the cascade and whether the summarizer's identity moderates downstream transmission:

3. **RQ3**: How does LLM-introduced bias propagate across human–LLM interaction chains?
4. **RQ4**: Does the source of summarization (human vs. LLM) influence the transmission of bias to downstream participants?

The primary, design-driven analyses for RQ1, RQ3, and RQ4 were prespecified; the RQ2 echo-chamber analysis was exploratory because alignment depended on participants' initial moral positions rather than experimental assignment.

3. Experiment 1: direct influence of LLM moral bias

3.1. Experiment task

Adapted from the Moral Machine framework (Awad et al., 2018), which extends the classic Trolley Problem to autonomous driving contexts, participants completed a moral decision-making task involving an autonomous vehicle (AV). In each scenario, the AV faced a moral dilemma: whether to continue its current path and crash into pedestrians or to swerve into a barrier, sacrificing its passengers. The AV's active decision to minimize total harm (swerving to save more lives) represented a utilitarian choice, whereas maintaining its course to avoid direct harm to passengers represented a deontological choice. The Moral Machine platform enabled the construction of realistic, ethically nuanced scenarios. Drawing on its dataset and design guidelines (Awad et al., 2018), we avoided highly polarizing contrasts (e.g., young vs. old, humans vs. animals, or large vs. small groups). Three scenarios were created using the Moral Machine custom scenario generator, each conceptually parallel to the Trolley Problem, and randomly assigned to participants. In all cases, the AV's inaction resulted in the deaths of pedestrians, and the number of pedestrians (N) exceeded the number of AV passengers ($N - 1$). This design minimized the salience of an overtly utilitarian preference by avoiding extreme moral asymmetries.

3.2. Experiment design

Experiment 1 used a 2 (Time: pre-vs. post-discussion; within-subjects) × 3 (LLM Bias: utilitarian, deontological, balanced; between-subjects) mixed design. For the **LLM Bias** condition, all prompts were pretested for perceived bias direction, emotional tone, and rhetorical strength (see Appendix A). All conversational responses were generated by Google's Gemini 2.5 Flash (see Appendix A). Across conditions, responses were standardized in length (~50 words) and style. Participants interacted with an LLM that consistently framed its responses according to one of three moral perspectives:

- **Utilitarian**: The LLM emphasized outcome-based reasoning, such as minimizing total harm or saving the greatest number of lives. Prompts guided the LLM to persuasively advocate for consequentialist moral principles.
- **Deontological**: The LLM advocated for rule- or duty-based ethics, such as not intentionally harming an individual regardless of consequences. Responses were framed to underscore individual rights and moral absolutes.
- **Balanced**: Serving as a baseline, the LLM presented a neutral and balanced perspective, offering arguments for both sides of the dilemma with equal rhetorical and emotional weight.

To confirm that the three prompts produced the intended moral framing, we analyzed all conversations using a keyword lexicon of utilitarian and deontological terms. The turn-level framing score (utilitarian minus deontological keyword frequency, averaged within participant) ordered exactly as designed ($M_{utilitarian} = 0.59$, $M_{balanced} = -0.22$, $M_{deontological} = -1.09$), $F(2, 117) = 150.95$, $p < .001$, all pairwise ds ≥ 1.77 (see Appendix Figure A1).

3.3. Measures

Participants completed a structured set of instruments assessing demographics and moral decision-making. Participants then completed a moral dilemma adapted from the classic Trolley Problem as described in Section 3.1. Participants' moral decisions were captured before and after the LLM conversation. At each time point, participants selected one of two options in an autonomous vehicle moral dilemma, either a utilitarian choice (sacrificing the passengers to save a larger number of pedestrians) or a deontological choice (not sacrificing). After each decision, they reported their confidence on a 7-point Likert scale (1 = not at all confident; 7 = extremely confident). To capture both direction and strength of participants' moral stance, we computed an *Opinion Strength Score* at both pre- and post-discussion time points. This composite measure assigns positive values to utilitarian choices and negative values to deontological choices, scaled by confidence.

3.3.1. Opinion Strength Score

To capture both the direction and certainty of participants' moral positions, we constructed an opinion-strength measure that combined decision choice and confidence. Let $c \in [1, 7]$ denote the participant's self-reported confidence rating, and let the moral choice be defined as $d \in \{U, D\}$, where U represents a utilitarian decision and D represents a deontological decision. We define the opinion strength score $s \in [-7, +7]$ as: $s = \begin{cases} +c, if\ d = U \\ -c, if\ d = D \end{cases}$.

3.4. Procedure

After providing informed consent and completing a commitment agreement to respond truthfully, participants provided demographic information and proceeded to the experimental task shown in Fig. 1. The task began with participants' initial evaluation of a Trolley Problem–style autonomous vehicle scenario (Section 3.1). Participants were randomly assigned to one of three LLM Bias conditions (utilitarian, deontological, balanced) and completed the following steps:

5. **Step 1**: Participants viewed a Moral Machine dilemma with a detailed explanation, made an initial decision, and reported confidence on a 7-point Likert scale.
6. **Step 2**: Participants engaged in a structured, text-based dialogue with an LLM adopting one of three stances (utilitarian, deontological, or balanced). The LLM responded using pretested prompt templates (Appendix A). Participants provided at least five responses, each at least 100 characters.
7. **Step 3**: Participants wrote a summary of the conversation (minimum 250 characters).
8. **Step 4**: Participants re-evaluated the same dilemma and reported their decision and confidence again.

Upon completing all steps, participants were fully debriefed. The debrief disclosed that the LLM had been assigned a specific moral orientation (utilitarian, deontological, or balanced) via a system prompt, and that this orientation was covertly maintained throughout the conversation. Participants were informed of the study's purpose and given the opportunity to withdraw their data before submission.

3.5. Participants

We recruited participants through the Prolific platform, an online participant-recruitment service built for research that provides a vetted, demographically diverse participant pool with identity verification and built-in pre-screening and data-quality controls. Eligibility criteria included: (1) residing in the United States, (2) being at least 18 years of age, and (3) fluency in English. Participants were compensated $8 USD for approximately 30 min of participation. A priori power analysis was conducted using G*Power 3.1 to determine the required sample size to detect medium-sized effects (f = 0.25) with 80% power (α = .05). For the one-way between-subjects design with three levels of LLM Bias (utilitarian, deontological, balanced), the analysis indicated a required sample size of N = 120. A total of 121 participants were recruited for this study. One participant was excluded for failing the attention check embedded in the personality questionnaire, resulting in a final sample of N = 120 participants, with a mean age of 38.6 (SD = 10.4) and 40.8% identified as female. Participants were evenly distributed across the three experimental conditions: deontological bias (n = 40), balanced (n = 40), and utilitarian bias (n = 40). The Georgia Institute of Technology Institutional Review Board determined this study to be exempt (IRB2025-624) on June 13, 2025. All participants provided informed consent before participation.

3.6. Results

3.6.1. Binary decision reversals were rare, yet confidence increased regardless of moral orientation

Prior to the AI discussion, participants showed a strong preference for utilitarian choices: 109 participants (90.8%) chose to hit the barrier to minimize casualties, and 11 (9.2%) chose the deontological option (i.e., maintain course rather than actively sacrifice the passenger). After discussion with LLMs, decision reversals were relatively rare. Across all conditions, only 14 participants (11.7%) changed their decision: balanced, 12.5% (5/40, 2 transitions from U→ D, and 3 transitions from D→U); deontological (D), 15.0% (6/40, all transitions from U→ D); and

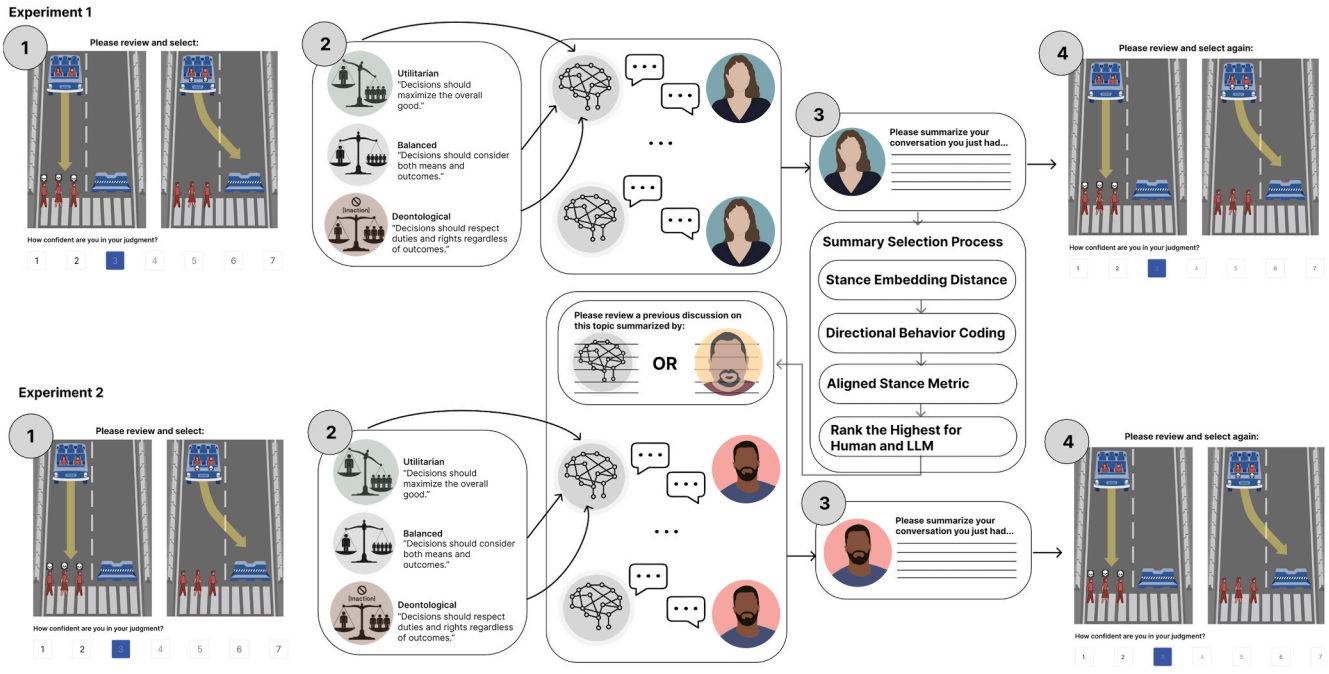

Fig. 1. Study flow for Experiments 1 and 2
Note. Summaries from Experiment 1 underwent a selection process (Section 4.1); selected summaries served as inputs for Experiment 2.

utilitarian (U), 7.5% (3/40, all transitions from U→ D). Addressing **RQ1**, a chi-square test revealed no significant association between bias condition and decision-change rate, $p > .05$, indicating no significant condition difference in the overall probability of any reversal, consistent with strong and stable pre-existing moral preferences.

Because the omnibus test collapsed across reversal direction, we conducted an exploratory directional analysis. Among the 109 participants who initially chose the utilitarian option, 6/37 (16.2%) in the deontological condition switched to the deontological option, compared with 2/36 (5.6%) in the balanced condition and 0/36 in the utilitarian condition, Fisher–Freeman–Halton exact $p = .026$. Of the 11 initially deontological participants, six switched to the utilitarian option, but these cells were too small for reliable comparisons. Thus, reversals were rare but directionally consistent with the assigned bias in both biased conditions.

Next, we tested whether AI bias influenced participants' confidence using a linear mixed-effects model (LMM) with fixed effects of Time and LLM Bias, participant-specific random intercepts and random slopes for Time, and restricted maximum likelihood (REML) estimation. As shown in Fig. 2A, the discussion significantly increased participants' confidence, $\beta = 0.58$, $z = 2.86$, $p = .004$, $d = 0.40$, 95% CI [0.18, 0.97], from pre-discussion ($M = 5.65$, $SE = 0.25$) to post-discussion ($M = 6.23$, $SE = 0.25$). Neither Bias condition nor Time × Bias interactions reached significance.

3.6.2. *Opinion strength shifted in the direction of LLM bias*

For opinion strength, as shown in Fig. 2B, results revealed a significant Time × Bias interaction for the deontological condition, $\beta = -2.23$, $z = -2.44$, $p = .015$, $d = -0.38$, 95% CI [−4.01, −0.44]. To interpret this interaction, we ran within-condition paired t-tests with Bonferroni correction ($\alpha = .05/3 = 0.0167$). In the balanced condition, as shown in Table 1, opinion strength showed minimal change from pre-discussion ($M = 4.90$) to post-discussion ($M = 5.48$), t (39) = 0.86, $p > .05$. In the utilitarian condition, opinion strength shifted toward utilitarian reasoning, from pre-discussion ($M = 4.40$) to post-discussion ($M = 6.10$), t (39) = 3.03, $p = .004$, $\Delta M = 1.70$, $d = 0.48$. In the deontological condition, opinion strength shifted toward deontological reasoning, from pre-discussion ($M = 5.03$) to post-discussion ($M = 3.38$), t (39) = −2.38, $p = .022$, $\Delta M = -1.65$, $d = -0.38$, which did not meet the Bonferroni-corrected threshold but indicates a meaningful small-to-medium effect.

Table 1
Simple effects analysis: Opinion strength changes within each condition.

Condition	M pre	M post	ΔM	t	df	*p*	d
Balanced	4.90	5.48	0.58	0.86	39	0.397	0.14
Deontological	5.03	3.38	−1.65	−2.38	39	**.022**†	−0.38
Utilitarian	4.40	6.10	1.70	3.03	39	**.004***	0.48

Note. N = 40 per condition. Tests are paired-samples t-tests comparing pre-to post-discussion opinion strength. Bonferroni-corrected α = .0167 (0.05/3 tests). Opinion strength ranges from −7 (strongly deontological) to +7 (strongly utilitarian). †p < .05 but does not meet Bonferroni-corrected threshold; effect size indicates meaningful shift. *p < .0167 (significant after Bonferroni correction)

Post-hoc tests at the post-discussion timepoint, with Bonferroni correction ($\alpha = .0167$), revealed no significant differences between balanced and deontological conditions (t (78) = 2.21, $p = .030$, $d = 0.50$) or between balanced and utilitarian conditions (t (78) = −1.03, $p > .05$, $d = -0.23$). However, the deontological and utilitarian conditions differed significantly, t (78) = −3.15, $p = .002$, $d = -0.70$, 95% CI [−1.16, −0.24], confirming divergence between the two biased framings by the end of the discussion. Overall, biased conditions shifted opinion strength in the predicted directions (deontological: $d = -0.38$; utilitarian: $d = 0.48$; Fig. 2B). Although people rarely reversed decisions, their confidence-weighted stance shifted toward the framing suggested by the LLM.

3.6.3. *Echo chamber analysis: alignment reinforces opinion strength and confidence*

Addressing RQ2, we examined whether alignment between a participant's initial moral position and the LLM's bias produced echo-chamber reinforcement. The balanced condition (n = 40) was excluded, and the remaining n = 80 participants were categorized as aligned (n = 39) when their pre-discussion choice matched the AI's moral orientation, or misaligned (n = 41) when it conflicted. As shown in Table 2, the sample was strongly skewed toward utilitarian starting

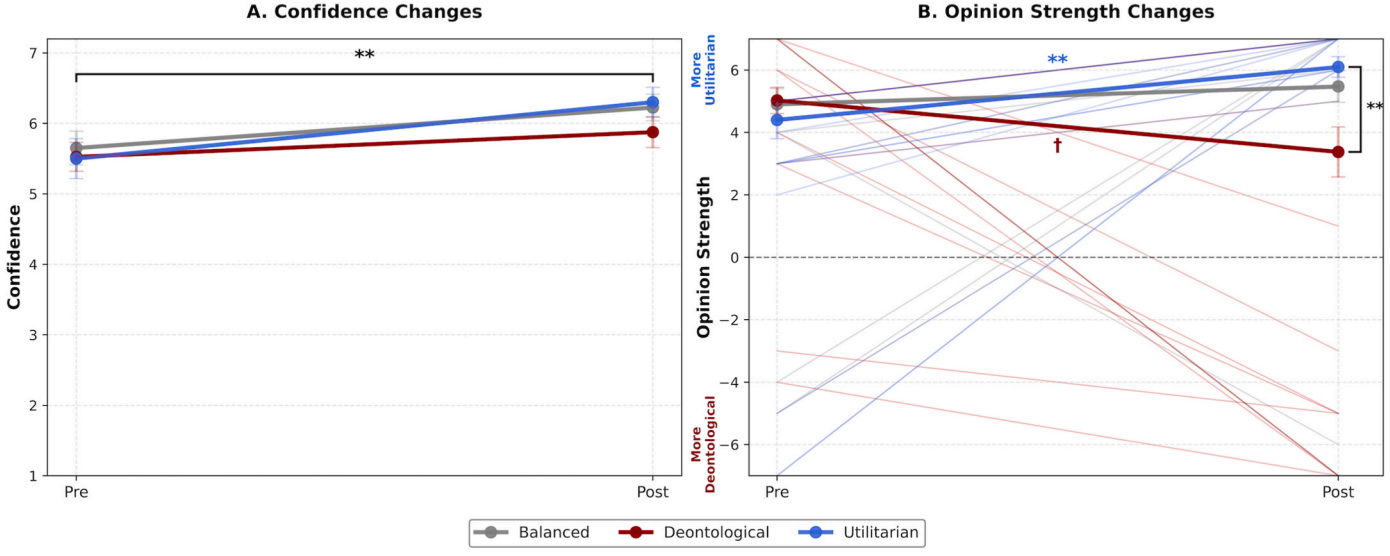

Fig. 2. Confidence and opinion strength changes by AI bias condition
Note. **Panel A** shows confidence increased uniformly across all conditions from pre-to post-discussion (p = .004, d = 0.40). **Panel B** displays opinion strength changes (−7 = deontological, +7 = utilitarian), with significance markers indicating within-condition effects: utilitarian-biased AI produced significant shifts toward utilitarian reasoning (**p = .004, d = .48), deontological-biased AI showed marginal shifts toward deontological reasoning (†, p = .022, d = −.38), and balanced condition remained stable. The two biased conditions diverged significantly by post-discussion (vertical bracket: **p = .002, d = −0.70). Individual trajectories shown for participants with $|\Delta| \geq 2$. Error bars: ± 1 SE. **p < .01; †p < .05 (marginal after Bonferroni correction).

Table 2
Composition and descriptive statistics by alignment status.

	Aligned (n = 39)	Misaligned (n = 41)
Utilitarian starters	36 (92.3%)	37 (90.2%)
Deontological starters	3 (7.7%)	4 (9.8%)

positions (73/80, 91.2%); consequently, the aligned group consisted almost entirely of utilitarian participants paired with utilitarian-biased AI (36/39), while the misaligned group consisted almost entirely of utilitarian participants who encountered deontological-biased AI (37/41). Because alignment status was therefore structurally confounded with initial moral position, we base our primary inference on the contrast within the 73 utilitarian starters, for whom "aligned" reduces to receiving utilitarian-biased AI and "misaligned" to receiving deontological-biased AI, holding initial position constant by design (Fig. 3).

Alignment reinforces opinion strength and confidence. We base inference on the 73 utilitarian starters, holding initial moral position fixed by design. Within the aligned group, participants became both more strongly utilitarian and more confident (opinion strength and confidence each rose from $M = 5.50$ to 6.33, $\Delta M = +0.83$, $t(35) = 3.61$, $p < .001$). The two outcomes differ, however, in how cleanly they separate aligned from misaligned participants. For opinion strength, the separation is robust: aligned participants strengthened their stance while misaligned participants moved in the opposite direction ($\Delta M = -1.62$, $p = .037$), yielding a significant Time × aligned-AI interaction ($\beta = 2.46$, $z = 3.10$, $p = .002$). For confidence the separation is weaker: misaligned participants also drifted slightly upward ($\Delta M = +0.22$), leaving the aligned-versus-misaligned confidence interaction only marginal ($\beta = 0.62$, $z = 1.90$, $p = .058$). The aligned confidence gain is not a ceiling artifact: although 17 of the 36 aligned starters were already at the scale maximum, the 19 below-ceiling starters still rose substantially ($\Delta M = +1.63$, $p < .001$). The full sample (aligned n = 39) gave a nominally larger confidence interaction ($\beta = 0.68$, $z = 2.23$, $p = .026$).

Confidence is decoupled from choice. Binary reversals were rare and strikingly one-sided: no aligned participant reversed their initial decision (0/36 within utilitarian starters; 0/39 in the full aligned sample), whereas misaligned participants occasionally did (6/37; 9/41 in the full sample; Fisher's exact $p = .002$). Critically, these reversers did not switch because their confidence collapsed: they crossed to the opposing choice at preserved or higher certainty (mean confidence change +0.17), so their large negative opinion-strength swing reflects a change in the sign of their stance rather than a loss of strength. Counter-attitudinal AI thus shifts opinion strength across the moral poles without denting how certain people feel, while aligned AI reinforces the existing stance and confidence. In both cases, confidence is preserved or amplified even as choices occasionally move.

3.6.4. *Certainty analysis of Human-LLM conversations*

To examine how certainty evolves during human–LLM dialogue, we analyzed linguistic markers of certainty across conversation turns. Certainty was operationalized using a weighted lexicon-based metric capturing explicit expressions of confidence and uncertainty adapted and extended from Panicek and Hricak (2016), comprising 56 words across four categories: high certainty (+2; 18 words), medium certainty (+1; 13 words), medium uncertainty (−1; 13 words), and high uncertainty (−2; 12 words) (see Appendix C). Raw scores were normalized by word count and transformed using tanh (score × 10) to bound values in [−1, +1]. We fitted a linear mixed-effects model predicting certainty scores with fixed effects of message index (conversation turn), speaker (user vs. model), bias condition (balanced, deontological, utilitarian), and their interactions, along with participant-level random intercepts. Table 3 presents the full model results.

Users systematically expressed higher certainty than AI across all conditions. The speaker effect was the dominant predictor in the model ($\beta = 0.42$, $p < .001$), where the users consistently expressed certainty scores approximately 0.13–0.34 units above the AI depending on the condition. This gap persisted regardless of ethical framing, suggesting a robust asymmetry in how humans and AI models linguistically signal confidence in moral deliberation.

Ethical framing modulated the AI's certainty level but not users'. While model certainty varied substantially across conditions (balanced vs. deontological: $d = -0.93$; balanced vs. utilitarian: $d = -0.48$), user certainty remained stable across all three framings (all $|d| < 0.13$, all $p > .30$). This dissociation indicates that the AI was responsive to its own instructed framing in ways that were lexically detectable, whereas users' expressed certainty was insensitive to condition at the level of overall means.

In biased conditions, the AI became progressively less certain over turns while users became slightly more certain, consistent with a framing-induced convergence pattern. As shown in Fig. 4, the three-way interactions were significant for the deontological condition ($\beta = 0.034$, $p = .017$) and the utilitarian condition ($\beta = 0.040$, $p = .005$). Simple slope analyses clarified these dynamics. For users, certainty decreased over turns in the balanced condition (slope = −0.019) but increased slightly over turns in both the deontological and utilitarian

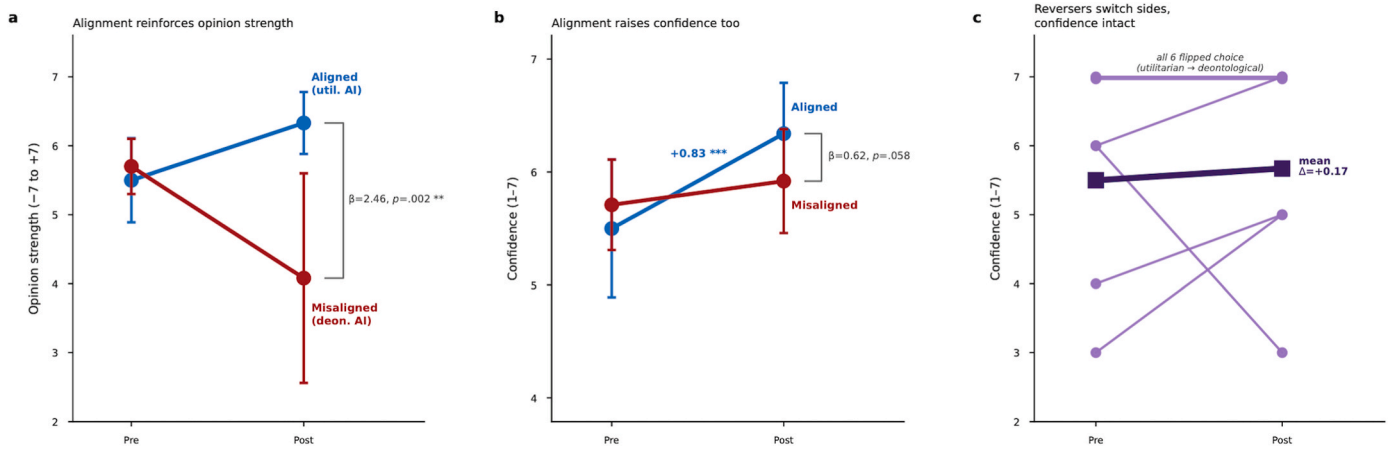

Fig. 3. The echo-chamber effect within utilitarian starters, holding initial moral position fixed
Note. Sample is the 73 utilitarian starters; aligned = utilitarian-biased AI (n = 36), misaligned = deontological-biased AI (n = 37). Points are group means ± 95% CI. (A) Opinion strength (−7 = deontological, +7 = utilitarian). (B) Confidence. (C) The six misaligned participants who reversed their choice did so at preserved confidence (mean Δ = +0.17), so the negative opinion-strength swing in (A) reflects a change in the sign of the stance.

Table 3
Linear mixed-effects model predicting certainty scores.

Predictor	β	SE	z	p	LL	UL
Main Effects						
Intercept	−0.012	0.042	−0.29	0.770	−0.10	0.07
Speaker (User)	0.422	0.062	6.80	**< .001**	0.30	0.54
Bias (Deontological)	0.263	0.060	4.38	**< .001**	0.15	0.38
Bias (Utilitarian)	0.193	0.060	3.22	**.001**	0.08	0.31
Message Index	0.004	0.007	0.51	0.609	−0.01	0.02
Two-Way Interactions						
Speaker × Deontological	−0.355	0.088	−4.04	**< .001**	−0.53	−0.18
Speaker × Utilitarian	−0.330	0.088	−3.75	**< .001**	−0.50	−0.16
Message Index × Speaker	−0.022	0.010	−2.22	**.026**	−0.04	−0.00
Message Index × Deontological	−0.012	0.010	−1.16	0.245	−0.03	0.01
Message Index × Utilitarian	−0.017	0.010	−1.71	0.087	−0.04	0.00
Three-Way Interactions						
Message Index × Speaker × Deontological	0.034	0.014	2.39	**.017**	0.01	0.06
Message Index × Speaker × Utilitarian	0.040	0.014	2.81	**.005**	0.01	0.07
Random Effects						
σ^2 participant	0.006	0.007				
σ^2 residual	0.080					

conditions (slope = 0.004 in each). For the model, certainty showed minimal change in the balanced condition (slope = 0.004) but declined across turns in the deontological (slope = −0.008) and utilitarian conditions (slope = −0.013). This divergence suggests that when the AI adopts an ethical framing stance, users progressively anchor to it, sustaining or gaining certainty as the conversation unfolds, while the model itself hedges more as the dialogue develops, pointing to a within-conversation influence dynamic driven by the AI's framing.

3.7. Experiment 1 discussion

Experiment 1 investigated whether interaction with a morally framed LLM altered moral decision-making without social transmission. The results converged on three patterns: a single conversation rarely overturned binary choices, opinion strength shifted toward the LLM's framing, and confidence increased across bias conditions. Expressed certainty also changed across conversational turns.

3.7.1. Stability of moral decisions under LLM influence

Participants' binary moral choices were highly stable following LLM interaction. Reversals were rare and statistically unrelated to the model's bias condition. This is consistent with the well-established literature on moralized attitudes, which are characterized by resistance to persuasion, temporal durability, and independence from conventional influence attempts (Luttrell et al., 2016; Skitka et al., 2005, 2021). When attitudes acquire a perceived moral basis, they become deeply embedded in identity and values, rendering them less susceptible to direct attitude-change attempts (Cheng et al., 2026; Luttrell et al., 2016). Crucially, individuals holding moralized attitudes often respond to persuasive pressure with reactance, intensifying rather than revising their position (Brehm, 1956; Skitka et al., 2021). This pattern also converges with research on moral dilemma paradigms, where utilitarian and deontological orientations are notably resistant to single-exposure manipulation, reflecting deeply held intuitions that respond more readily to framing cues than to overt persuasion (Luttrell et al., 2016). These results caution against interpretations that brief LLM interactions will rapidly homogenize moral preferences: direct decision change appears to require more sustained or socially embedded influence processes.

3.7.2. Confidence increased across bias conditions in experiment 1

In Experiment 1, categorical decisions rarely changed, but confidence increased after LLM interaction across all three bias conditions, with no reliable differences among them. This pattern was consistent across the three moral orientations tested in this experiment. Because the study did not directly test explanatory mechanisms, two processes

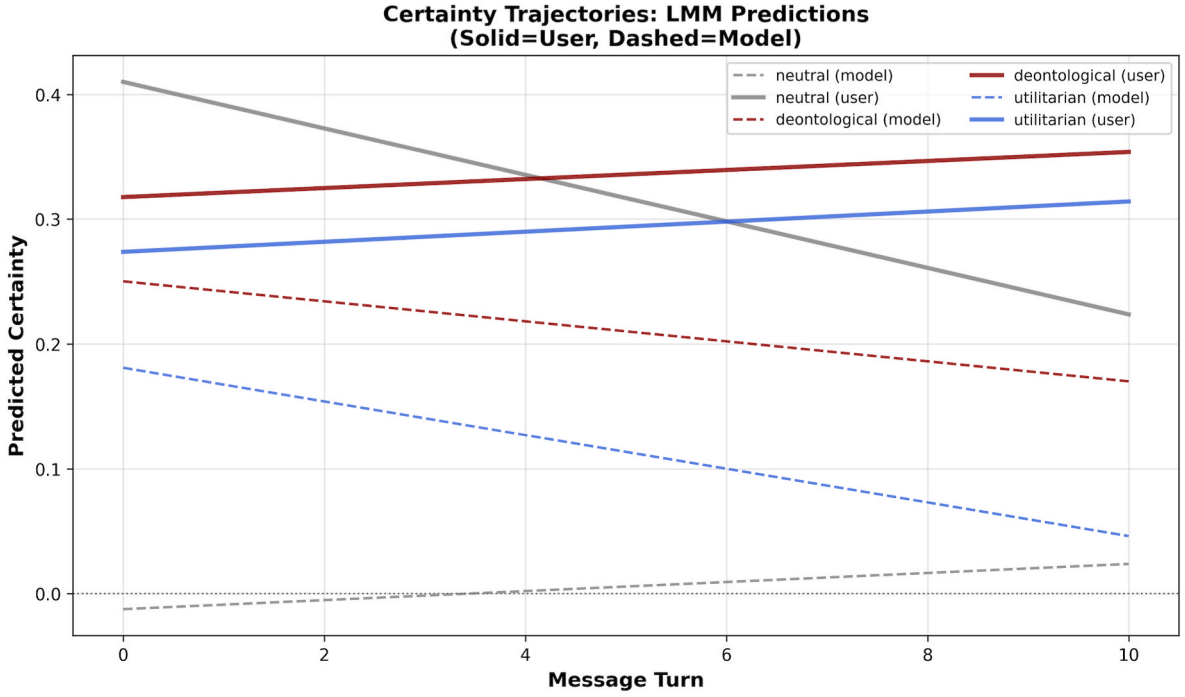

Fig. 4. Predicted certainty score trajectories across conversation turns by bias condition and speaker.
Note. Users in biased conditions (solid lines) showed slight increases in expressed certainty, whereas model certainty declined.

offer tentative interpretations. First, sycophantic tendencies documented in Reinforcement Learning from Human Feedback (RLHF)-trained models can produce affirming feedback that validates users' views, potentially increasing certainty even when such certainty is not epistemically warranted (Batista & Griffiths, 2026). Second, the structured dialogue may have fostered self-validation. Participants were required to articulate and summarize the reasons underlying their initial moral judgments, making those reasons more accessible for metacognitive evaluation. Under the self-validation hypothesis, people rely more heavily on thoughts they perceive as valid: confidence in favorable thoughts facilitates persuasion, whereas confidence in counterarguments facilitates resistance (Barden & Petty, 2008; Petty et al., 2002). Repeated articulation and responsive feedback may therefore have increased the perceived validity of participants' reasoning regardless of its moral direction. These findings raise the possibility that human–LLM dialogue can leave users more certain of their moral judgments without necessarily improving their quality or changing their decisions.

3.7.3. Alignment reinforces prior views, whereas misalignment shifts direction without reducing confidence

Because the sample was overwhelmingly utilitarian at baseline, the primary echo-chamber analysis was restricted to utilitarian starters, thereby holding initial moral choice constant. Within this comparison, alignment affected opinion direction more clearly than confidence. An aligned, utilitarian-biased LLM strengthened participants' utilitarian opinion strength and confidence, whereas a misaligned, deontological-biased LLM shifted opinion strength toward deontology without reliably reducing confidence. Correspondingly, the Time × Alignment interaction was significant for opinion strength but only marginal for confidence. Even the misaligned participants who reversed their decisions crossed to the opposing position with preserved or slightly increased confidence. Alignment therefore determined the direction of movement more strongly than whether confidence was maintained.

These patterns suggest several possible pathways. In aligned exchanges, stance-consistent responses may validate favorable thoughts and reinforce the existing position. Among misaligned participants who maintained their choice, resisting a counter-attitudinal LLM may have increased certainty by signaling that their original position was defensible. Tormala and Petty (2002) similarly found that perceived successful resistance to a strong persuasive challenge increased attitude certainty even without attitude change. Participants who reversed, however, became more aligned with the LLM without losing confidence. This pattern is compatible with the amplification hypothesis, under which certainty strengthens relatively settled attitudes but can coexist with thoughtful change when attitudes are ambivalent (Clarkson et al., 2008). Overall, alignment shaped opinion direction, but confidence persisted whether participants were reinforced or redirected.

3.7.4. Linguistic certainty patterns across conversation turns

The turn-by-turn linguistic analyses described how expressed certainty changed during the conversations. Participants consistently expressed higher certainty than the model, suggesting human utterances differ structurally from AI utterances in moral debates. More importantly, certainty trajectories diverged across conditions: humans in neutral conversations exhibited declining certainty across turns, while users in value-biased conversations maintained or modestly increased certainty. Concurrently, the model expressed higher certainty in biased conditions, but its certainty declined across turns regardless of condition. This bidirectional pattern is consistent with the possibility that value-consistent framing helped maintain users' expressed certainty, but it does not establish that framing caused the post-discussion confidence increase (Batista & Griffiths, 2026). One speculative interpretation is algorithmic authority: users may attribute epistemic legitimacy to LLM outputs when the model's framing appears to validate their prior beliefs (Batista & Griffiths, 2026).

4. Experiment 2: social influence of bias in Human-LLM cascades

Experiment 1 demonstrated that LLM moral bias can systematically shift human confidence in moral decision-making following a single human–LLM interaction. However, real-world moral judgments are rarely formed in isolation. Instead, individuals are often exposed to others' opinions through reposts, summaries, and curated discussions, creating social influence dynamics and information cascades. Thus, Experiment 2 examined how moral bias introduced by LLMs propagates across individuals when prior interactions are summarized and transmitted downstream. Experiment 2 employed a 2 (**Time**: pre-vs. post-discussion; within) × 3 (**LLM Bias**: utilitarian, deontological, balanced; between) × 2 (**Summarizer**: human, LLM; between) factorial design. This design focuses on how LLM-induced bias is retained, attenuated, or amplified through (a) prior summarization within a sequential decision-making task and (b) the summarization source. The LLM-bias manipulation remained the same as in Experiment 1. Before interacting with the LLM, participants received an additional summary of the prior discussion generated either by a human or by the LLM. This manipulation allowed us to examine whether human-generated summaries attenuate or amplify the transmission of bias in the cascade compared to summaries created by the biased LLM itself.

4.1. Summary selection

Experiment 2 required a principled basis for choosing which Experiment 1 conversations to pass downstream. We selected summaries from conversations in which participants shifted toward the LLM's moral framing. These summaries, written by participants or generated by the LLM, served as the inputs downstream participants read. This kept the cascade manipulation consistent across subjects and grounded it in validated cases of moral influence rather than arbitrary discourse. The strategy models a high-risk cascade, testing whether strongly aligned moral framings, once formed, survive summarization. It also operationalizes an online echo chamber, where people mainly encounter stance-consistent content (Cinelli et al., 2021). To identify these conversations, we developed a directionally aware stance-shift metric that combines semantic change with behavioral alignment.

First, we quantified the magnitude of conversational stance change using sentence embeddings (i.e., all-MiniLM-L6-v2). For each participant in Experiment 1, we computed the cosine distance between their first and final conversational messages with the LLM, yielding a continuous index (0–1) of semantic movement across the dialogue. However, semantic change alone is insufficient to distinguish persuasion from resistance. A participant who increasingly argues against the LLM's position may exhibit large semantic change despite rejecting the AI's framing. To address this ambiguity, we incorporated directional alignment based on participants' pre- and post-discussion moral decisions relative to the LLM's assigned bias. Specifically, in utilitarian-biased conversations, choosing to hit the barrier (minimizing total casualties) was coded as aligned, whereas in deontological-biased conversations, choosing to hit pedestrians (avoiding intentional harm) was coded as aligned. This behavioral coding allowed us to determine whether semantic movement occurred toward or away from the AI's intended moral direction. The resulting aligned stance-shift metric captures both the extent and direction of influence, identifying conversations in which participants meaningfully converged toward the LLM's moral framing. Next, for each bias condition, we ranked conversations based on this metric and selected summaries corresponding to highly aligned cases (see Appendix B).

4.2. Procedure

As shown in Fig. 1, the experimental procedure was identical to that of Experiment 1 (Section 3.4), with one key difference: participants viewed a selected summary prior to the discussion with the LLM. All

summaries were selected using the aligned stance-shift metric described in Section 4.1, which identified interactions that exhibited the strongest convergence toward the assigned LLM moral bias. The author of the summary (human vs. LLM) was randomly assigned, but the stance of the summary aligned with the current LLM bias condition (utilitarian, deontological, or balanced). Participants reviewed the summaries for 30 s before they engaged in a text-based discussion with the LLM, providing at least five responses in a 10-round conversation with the LLM agent. Summaries were available for participants to refer to during the conversation phase. Participants then re-evaluated the same moral dilemma and reported their moral decision and confidence. Finally, participants completed the same post-task measures used in Experiment 1. Debriefing followed the same protocol as Experiment 1, disclosing the LLM's covert moral orientation and the summary selection procedure.

4.3. Results

Participant recruitment criteria were identical to those of Experiment 1. A separate a priori power analysis for Experiment 2's 3 (LLM Bias) × 2 (Time) × 2 (Summarizer) repeated-measures MANOVA design was conducted in G*Power 3.1 (F tests; f = 0.25, α = .05, power = 0.80). This indicated a required sample size of N = 114. The recruited sample of n = 120 for Experiment 2 therefore exceeds this threshold and is sufficient for the primary analyses. The combined sample across both experiments comprised N = 240 participants (Experiment 1: n = 120; Experiment 2: n = 120). Similar to Experiment 1, LMMs were used with fixed effects of LLM Bias, Time, and Summarizer, with participant-specific random intercepts and random slopes for time, estimated via restricted maximum likelihood (REML).

Pre-discussion opinion strength did not differ between Experiment 1 ($M = 4.78$, $SD = 3.24$) and Experiment 2 ($M = 4.47$, $SD = 3.56$, $p > .05$), confirming successful randomization. Decision change rates were comparable across experiments: 11.7% (14/120) in Experiment 1 and 15.0% (18/120) in Experiment 2, $p > .05$. In Experiment 2, among the 109 initially utilitarian participants, 8/36 (22.2%) in the deontological condition switched to the deontological option, compared with 2/37 (5.4%) in the balanced condition and 3/36 (8.3%) in the utilitarian condition; these rates did not differ significantly ($p > .05$). Only 11 participants initially chose the deontological option, precluding reliable directional comparisons. Thus, social transmission did not produce a statistically reliable increase in binary decision reversals.

4.3.1. Confidence increased across bias conditions in both experiments

To address **RQ3**, we again found a significant main effect of Time, b = 0.58, SE = 0.20, z = 2.85, p = .004, replicating the confidence increase observed in Experiment 1. Neither the Time × Bias interaction nor the three-way Time × Bias × Experiment interaction reached significance, indicating that confidence gains were uniform across bias conditions and did not vary by cascade position. Follow-up paired t-tests (Bonferroni-corrected α = .017) confirmed significant pre-to-post confidence increases in all three conditions when data from the two experiments were combined (Table 4). The utilitarian condition yielded the largest increase (d = 0.63), followed by balanced (d = 0.50) and deontological (d = 0.29) conditions. Within these experiments, confidence increased across the tested bias conditions and both cascade positions.

Table 4
Simple effects for confidence by bias condition, combined across experiments 1 and 2.

Condition	M pre	M post	ΔM	SD Δ	t (79)	p	d
Balanced (n = 80)	5.65	6.19	0.54	1.08	4.46	**< .001***	0.50
Deontological (n = 80)	5.50	5.95	0.45	1.53	2.62	**.010***	0.29
Utilitarian (n = 80)	5.42	6.15	0.72	1.15	5.65	**< .001***	0.63

4.3.2. Bias shifted opinion strength in condition-aligned directions, attenuated at the second cascade position

A significant Time × Bias interaction emerged for the deontological condition, b = −2.23, SE = 1.00, z = −2.23, p = .026, 95% CI [−4.18, −0.27], indicating that opinion strength changed differentially over time as a function of bias condition. The analysis of opinion strength revealed no significant main effects of Time, Bias, or Experiment (all ps > 0.05). Critically, the three-way Time × Bias × Experiment interactions were non-significant for both deontological (p > .05) and utilitarian (p > .05) conditions. Contrary to our primary hypothesis, bias effects were not amplified at the second cascade position; the persuasion pattern did not differ statistically between Experiment 1 (direct AI discussion) and Experiment 2 (summary-mediated exposure).

Follow-up paired t-tests (Bonferroni α = .017) decomposed the significant Time × Bias interaction. The balanced condition showed no significant change (d = 0.12, p > .05), confirming the control condition as a valid reference. The deontological condition produced a significant shift toward deontological positions (ΔM = −1.45, d = −0.27, p = .017), reaching the Bonferroni-corrected threshold. The utilitarian condition produced a shift toward utilitarian positions (ΔM = +1.07, d = 0.26, p = .023). This effect was significant at the uncorrected 0.05 level but did not survive Bonferroni correction. Thus, the deontological shift constitutes the stronger and more reliable evidence of AI-induced persuasion, while the utilitarian result warrants cautious interpretation.

Although the three-way interaction did not reach significance, effect size comparisons across experiments revealed a theoretically meaningful descriptive pattern (Table 5). The utilitarian bias effect attenuated substantially from Experiment 1 to Experiment 2 (d = +0.48 vs. d = +0.10; Δd = −0.38), a reduction of approximately 80% across the cascade. The deontological effect showed more modest attenuation (d = −0.38 vs. d = −0.20; Δd = +0.17), retaining roughly half of its magnitude at the second cascade position. The balanced condition remained near zero at both positions (Δd = −0.05). Taken together, these effect sizes show a descriptive attenuation pattern, especially for utilitarian persuasion, but the non-significant three-way interaction does not support a statistically reliable difference in bias effects between experiments.

4.3.3. AI-generated summaries produced stronger deontological shifts than human-generated summaries

To address **RQ4**, a 3 (Bias) × 2 (Time) × 2 (Summarizer: AI, human) LMM was fit to Experiment 2 data (n = 120). As shown in Fig. 5, a significant three-way Time × Bias (deontological) × Summarizer (AI) interaction emerged, b = −4.90, SE = 2.12, z = −2.31, $p = .021$, 95% CI [−9.05, −0.75]. The corresponding interaction for utilitarian content was not significant (p > .05). At post-discussion, participants in the deontological condition who received an AI-generated summary reported significantly lower opinion strength ($M = 1.00$, $SD = 6.10$) than those who received a human-generated summary ($M = 5.05$, $SD = 3.85$), t (38) = −2.51, $p = .016$, $d = -0.80$. This indicates that AI summaries were associated with a substantially stronger shift toward the

Table 5
Effect Sizes (Cohen's d) for Opinion Strength Change by Experiment and Cascade Position.

Condition	Experiment 1 d	Experiment 2 d	Δd	Pattern
Balanced (n = 40)	0.14	0.09	−0.05	Similar
Deontological (n = 40)	−0.38	−0.20	0.17	Similar
Utilitarian (n = 40)	0.48	0.10	−0.38	Attenuation

Note. Positive d = shift toward utilitarian positions; negative d = shift toward deontological positions. |Δd| > 0.20 indicates a meaningful difference across cascade positions.

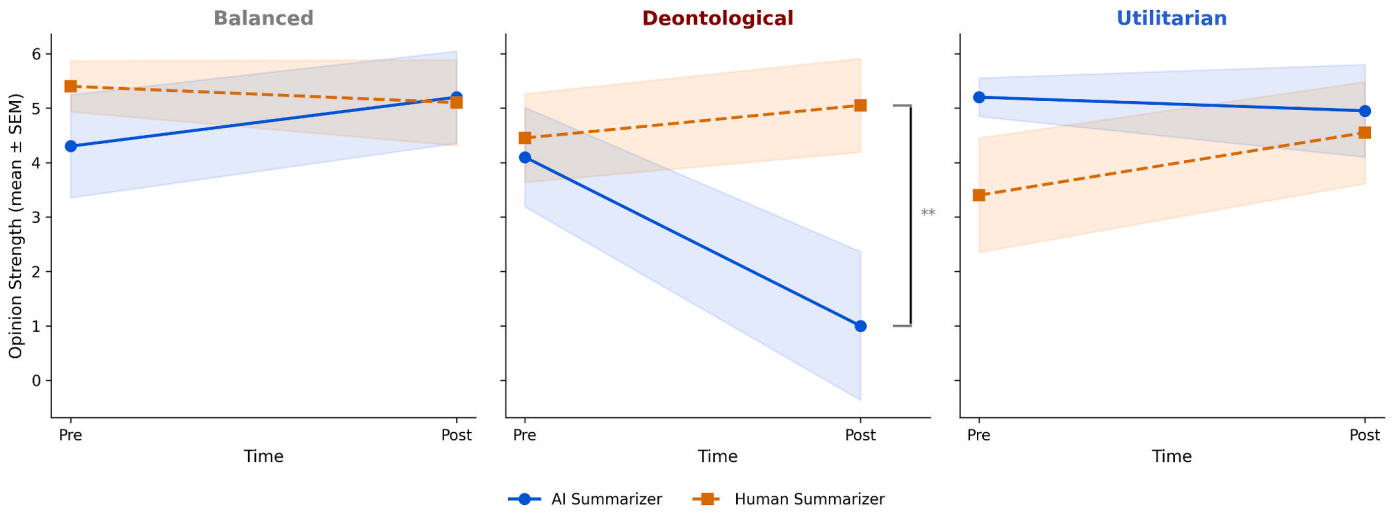

Fig. 5. Effect of summarizer identity on opinion strength grouped by bias condition.

deontological pole than human-authored summaries.

5. General discussion

5.1. LLMs as confidence anchors in moral decision-making

Across both experiments, the primary empirical pattern of LLM moral influence was confidence amplification rather than decision change. In direct human–LLM interaction (Experiment 1) and social cascade conditions (Experiment 2) alike, participants rarely reversed their binary moral choices, yet post-interaction confidence increased reliably. Simultaneously, opinion strength, a measure integrating decision direction and confidence level, shifted in the direction of the LLM's moral framing. In these experiments, LLM moral influence was more evident in the intensity than in the direction of moral commitment.

Categorical decision stability follows directly from the moralized attitudes literature: morally convicted attitudes are comparatively immune to authority and peer influence (Skitka et al., 2021). Our results extend that immunity to a new class of source, an LLM interlocutor, whose conversational and seemingly knowledgeable delivery might have been expected to overcome it. One difference from recent studies reporting that AI advice shifts moral judgments concerns the timing of the initial choice (Krügel et al., 2023, 2026). Those studies elicit a first-pass judgment about a dilemma the participant has not previously ruled on, whereas the present paradigm records a committed binary choice and confidence before LLM exposure. This distinction maps onto the free-choice paradigm (Brehm, 1956): once a person commits to one alternative over another, choice-induced attitude change follows, the chosen option is revalued upward and the rejected one downward, and dissonance reduction renders the decision progressively more resistant to reversal (Festinger, 1957). Because participants in our design commit to a moral choice before interacting with the LLM, confirmatory input may have bolstered the standing decision rather than overturned it. Prior commitment may therefore operate as a boundary condition on AI moral influence: absent a committed position, advice can move the choice itself; given one, choice-induced revaluation redirects the same influence onto the confidence with which the original choice is held.

The echo-chamber analysis (RQ2), restricted to the utilitarian starters so that alignment was not confounded with initial moral position, showed that alignment was more clearly associated with opinion direction than with confidence. Alignment separated opinion strength sharply: aligned participants strengthened their stance while misaligned participants moved toward the opposing pole. Confidence only marginally separated aligned and misaligned groups because misaligned participants also drifted upward. In other words, exposure to disagreement, the standard corrective for echo chambers, did not induce doubt here, but rather equally confident people holding a different position. Because the study did not directly test why confidence persisted, the mechanisms underlying this pattern remain uncertain. (Barden & Petty, 2008). In this sample, encountering disagreement shifted some users' positions without reducing their certainty.

5.2. Cascade attenuation rather than amplification

Experiment 2 tested whether LLM-induced moral bias would propagate and amplify across a social cascade (RQ3). The between-experiment interaction was not significant, so the data did not show that persuasion differed reliably between the two experiments. Descriptively, utilitarian effect sizes declined by approximately 80% from Experiment 1 to Experiment 2 ($d = +0.48$ to $d = +0.10$), while deontological effects showed more modest attenuation, retaining roughly half their magnitude ($d = -0.38$ to $d = -0.20$). Confidence gains replicated uniformly across cascade positions, independent of bias content. The descriptive attenuation pattern should be interpreted in light of Experiment 2's high-risk design, which used highly aligned summaries. One speculative interpretation follows the amplification hypothesis: raising certainty strengthens univalent one-sided attitudes but weakens ambivalent mixed ones (Clarkson et al., 2008). Second-position participants may have held more ambivalent attitudes because they inherited a summarized position rather than developing the argument themselves.

The attenuation findings contrast with the amplifying feedback loops documented by Glickman and Sharot (2025) and with the larger effects of social influence in people's accuracy in detecting misinformation (Lu et al., 2022). Two features of this cascade may help explain the descriptive attenuation pattern. First, the domain may matter. Lu et al. (2022) studied misinformation detection, a domain where participants face genuine factual uncertainty and treat AI credibility signals as informative. Moral judgments engage deeply held values and identity commitments that are structurally more resistant to external influence (Luttrell et al., 2016; Skitka et al., 2021). Participants at the second cascade position reported pre-existing moral orientations as stable as those at the first, which may have left less leverage for transmitted bias.

Second, the present cascade paradigm, in which social transmission occurs through a single written summary, is structurally different from the algorithmic feedback loops in which amplification has been observed. Echo-chamber amplification is often driven by repetition, network volume, and algorithmic curation, not by the fidelity of any single person-to-person retelling. Glickman and Sharot (2025) demonstrated amplification when an AI algorithm was iteratively trained on

biased human data, magnifying small biases through computational optimization. A single-summary cascade may therefore be more lossy than an iterative algorithmic feedback loop. Section 5.3 describes the observed differences between the selected human- and AI-generated summaries.

5.3. AI summarizers preserve moral bias more faithfully than human summarizers

Although overall bias attenuated across the cascade, the source of summarization moderated this attenuation (RQ4). Participants who received AI-generated summaries of deontological content shifted significantly toward deontological positions ($d = -0.43$), whereas those who received human-generated summaries of the same content showed no meaningful shift ($d = +0.14$), yielding a large cross-summarizer divergence ($\Delta d = -0.57$). AI-generated summaries preserved the explicit normative claim and transmitted deontological framing faithfully. In the selected examples, this pattern is consistent with greater preservation of the source's explicit normative claim. LLMs reproduce the explicit propositional content and argumentative structure of source material, preserving the moral framing, key claims, and certainty language of the original exchange.

In the selected examples, human-generated summaries appeared to preserve position more clearly than persuasive structure. Participants abstracted and reformulated the conversations in their own words, a pattern that may reflect filtering through existing schemas and salience judgments (Bartlett, 1932). Bartlett's serial reproduction studies demonstrated that argumentative content tends to be reduced to a "bald expression of conventional opinion" as it passes through successive human recall, stripping away the specific arguments and emotional register that carry rhetorical force. In these examples, positional content appeared to survive more clearly than persuasive scaffolding. This contrast is illustrated directly by the study's summaries in Appendix B: the AI-generated deontological summary explicitly states the moral argument ("emphasizing the impermissibility of intentionally harming innocent individuals"), whereas the human-generated counterpart focuses more on interpersonal experiences, omitting the core normative claim. When LLMs summarize morally biased content, they may therefore act as bias-preservation mechanisms, transmitting the original framing's rhetorical structure to downstream readers who may perceive the summary as neutral and objective (Lu et al., 2022).

5.4. Deontological effects were descriptively more resilient than utilitarian effects

Across the social cascade, deontological bias was descriptively more resilient than utilitarian bias. Deontological persuasion remained the only statistically credible signal of AI-induced opinion change at the second cascade position, whereas the utilitarian effect was small. The summarizer analysis showed a similar asymmetry: the source of summarization made no detectable difference for utilitarian content, yet produced a large divergence for deontological content. Two potential reasons may account for this asymmetry.

First, the strong utilitarian baseline (90.8% of participants chose the utilitarian option prior to any LLM interaction) reflects a well-documented tendency in switch-type moral dilemmas (Awad et al., 2018; Greene et al., 2001). This near-ceiling distribution limits additional opinion movement in the utilitarian direction at the second cascade position. Deontological arguments, by contrast, represented a minority position that most participants had not endorsed. One speculative account is minority influence: because deontological arguments were minority-consistent in this sample, their effects may differ from those of majority-consistent utilitarian arguments (Moscovici, 1980; Wood et al., 1994). This near-ceiling utilitarian rate is consistent with large cross-cultural evidence that a strong majority favors the utilitarian option in switch-type dilemmas (Awad et al., 2020). We note, however, that this asymmetry also yields a competing, non-motivational account of the resilience pattern: because so few participants began in the deontological position, there was simply more room to move toward deontology than toward the already-saturated utilitarian pole (differential headroom). The within-utilitarian-starter echo-chamber analysis reported above speaks against a pure-headroom explanation: aligned utilitarian participants, who had little room to move, still showed reliable confidence gains. But we cannot fully separate genuine differences in persuasive force from differential opportunity with the present distribution, and we flag this in the Limitations.

Second, deontological attitude shifts carry motivational staying power rooted in self-image protection. Individuals who make utilitarian decisions are perceived as less warm, trustworthy, empathetic, and moral than those who make deontological choices (Everett et al., 2018; Uhlmann et al., 2013). Because moral character exerts stronger influence than competence in social impression formation (Brambilla et al., 2011; Leach et al., 2007), people tend to align expressed moral preferences toward deontological positions to avoid the social costs of appearing utilitarian (Bostyn & Roets, 2017; Reynolds et al., 2019). In human-AI contexts, this pressure is further amplified: humans who endorse utilitarian positions face harsher social judgment than machines that do so (Malle et al., 2015; Chu & Liu, 2023). If this motivation operated here, it could help sustain a deontological shift and confer durability that utilitarian persuasion may lack. The same speculative account could also explain the AI summarizer difference reported above. AI-generated summaries may have preserved the explicit normative claim and thereby reactivated identity-protective motivation, whereas human summaries may have reduced this trigger. The study did not measure self-image protection, so this explanation requires direct testing.

5.5. Implications

These findings carry several practical implications for the design and governance of AI systems deployed in morally consequential contexts. First, LLMs that increase users' confidence without improving the quality of their moral reasoning risk producing systematic overconfidence among users who resist legitimate counterarguments and propagate their views with unwarranted certainty (Clarkson et al., 2008). Dialogue systems used in advisory or decision-support roles should consider making the LLM's normative orientation explicit to users, presenting opposing arguments with comparable elaboration, or incorporating reflection prompts that encourage users to examine the basis for their confidence rather than simply report its level (Buçinca et al., 2021; Petty et al., 2002).

Second, the AI summarizer finding points to a less-recognized vector of moral influence. Users who encounter AI-generated summaries of prior discussions may not recognize that the summary preserves a particular moral framing; they may instead perceive it as objective reportage. Platforms and tools that use LLMs to condense prior discourse should consider whether flagging the framing orientation of summarized content can mitigate inadvertent propagation of morally biased framings. Third, the overall patterns on stable decisions and inflated confidence suggest that, in similar settings, LLM moral influence may appear less as mass opinion conversion and more as an incremental promotion of existing moral positions across populations, progressively narrowing the space for deliberate moral reconsideration. This warrants consideration in AI governance frameworks (Gabriel et al., 2025) and in the deployment policies of organizations using LLMs in ethically consequential advisory roles.

5.6. Limitations and future directions

Several limitations should be considered. The study used a single LLM, a single dilemma domain (autonomous vehicle trolley problems), and a US English-speaking Prolific sample, limiting generalizability

across LLM models, moral domains, and cultural contexts. Second, the baseline moral distribution was highly skewed: 90.8% of participants began with the utilitarian choice, so the echo-chamber (RQ2) analysis rests almost entirely on utilitarian starters. A balanced test of the echo-chamber hypothesis will require a starting sample that is more evenly split between utilitarian and deontological positions through stratified sampling, so that aligned and misaligned conditions can be populated independently of initial moral position. In addition, the cascade was restricted to two positions: whether attenuation continues monotonically or eventually reverses in longer chains remains an open question. Relatedly, because Experiment 2 operationalized an echo chamber by design (Section 4.1), a natural extension is to treat summary selection as an independent variable, comparing high-alignment, random, and low-alignment summaries to chart how selection strength modulates downstream transmission. Future work should examine confidence propagation in longer cascades and in higher-stakes moral domains, including clinical triage, legal judgment, and hiring decisions. Intervention studies are particularly warranted: does disclosing an LLM's moral framing prior to dialogue reduce post-interaction confidence inflation? Do structured reflection prompts restore appropriate calibration after interaction?

6. Conclusion

Across these two experiments, LLM moral influence was more evident in confidence than in categorical decision reversal. Participants who conversed with a morally biased LLM rarely changed their categorical moral choice. Instead, they emerged more certain of their initial positions, with confidence-weighted opinion strength shifting in the direction of the model's framing. In the high-risk cascade design, bias-effect sizes were descriptively smaller in Experiment 2 than in Experiment 1, although the between-experiment interaction was not significant. Deontological framing was descriptively more resilient than utilitarian framing, and AI-generated summaries produced a stronger deontological shift than human-generated summaries, identifying the source of summarization as a distinct mechanism of bias transmission.

These findings identify questions for future research on the design and governance of conversational AI systems. Experiment 2 tested a high-risk but short cascade: downstream participants read highly aligned summaries selected from influential prior conversations and then interacted with the same biased LLM. Bias nevertheless did not amplify at the second position, suggesting that alignment-based selection alone may be insufficient when exposure is limited and persuasive content is compressed through summarization. Future studies should test whether longer chains, repeated exposure, or higher-fidelity summaries gradually harden moral positions, increase willingness to advocate publicly, or heighten susceptibility to downstream social influence. Such studies should also test whether transparency interventions that flag an LLM's bias indicator on AI-generated summaries restore appropriate uncertainty before consequential decisions. The present findings identify confidence as a distinct outcome that should be examined across models, settings, populations, and longer interaction sequences.

CRediT authorship contribution statement

Mengyao Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Trevor Patten:** Writing – review & editing, Validation, Data curation, Conceptualization. **Nishthaa Lekhi:** Writing – review & editing, Validation, Methodology, Conceptualization. **Areen Alsaid:** Writing – review & editing, Validation, Project administration, Methodology, Conceptualization.

Ethics declaration

Written informed consent to take part in the study and to publish the article has been obtained from all participants or their legal representatives. The privacy rights of participants have been observed.

This study was performed in compliance with relevant laws, regulatory frameworks and guidelines where the research took place. This study was approved by the Georgia Institute of Technology. (Approval No. IRB2025-624)

Declaration of generative AI

During preparation of this work, the authors used Claude (Anthropic) to assist with manuscript editing, language refinement, and code review, and Codex (OpenAI) to assist with manuscript editing and language refinement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank members of the Hybrid Intelligence at Georgia Tech for their discussions and comments. In particular, we thank Shaurya Kumar for contributing to the development of the experimental platform.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbr.2026.101473.

Data availability

Data are available from the corresponding author, Mengyao Li (mengyao.li@gatech.edu), upon request.

References

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus, 3* (10), Article pgae403. https://doi.org/10.1093/pnasnexus/pgae403

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

Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences, 117*(5), 2332–2337. https://doi.org/10.1073/pnas.1911517117

Barden, J., & Petty, R. E. (2008). The mere perception of elaboration creates attitude certainty: Exploring the thoughtfulness heuristic. *Journal of Personality and Social Psychology, 95*(3), 489–509. https://doi.org/10.1037/a0012559

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Batista, R. M., & Griffiths, T. L. (2026). A rational analysis of the effects of sycophantic AI. arXiv Preprint arXiv:2602.14270. https://doi.org/10.48550/arXiv.2602.14270

Bostyn, D. H., & Roets, A. (2017). An asymmetric moral conformity effect: Subjects conform to deontological but not consequentialist majorities. *Social Psychological and Personality Science, 8*(3), 323–330. https://doi.org/10.1177/1948550616671999

Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A., & Van Bavel, J. J. (2017). Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences, 114*(28), 7313–7318. https://doi.org/10.1073/pnas.1618923114

Brambilla, M., Rusconi, P., Sacchi, S., & Cherubini, P. (2011). Looking for honesty: The primary role of morality (vs. sociability and competence) in information gathering.

European Journal of Social Psychology, 41(2), 135–143. https://doi.org/10.1002/ejsp.744
Brehm, J. W. (1956). Postdecision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology, 52*(3), 384. https://doi.org/10.1037/h0041006
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
Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. *Science, 391*(6792), Article eaec8352. https://doi.org/10.1126/science.aec8352
Chu, Y., & Liu, P. (2023). Machines and humans in sacrificial moral dilemmas: Required similarly but judged differently? *Cognition, 239*, Article 105575. https://doi.org/10.1016/j.cognition.2023.105575
Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences, 118*(9), Article e2023301118. https://doi.org/10.1073/pnas.2023301118
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
Everett, J. A. C., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology, 79*, 200–216. https://doi.org/10.1016/j.jesp.2018.07.004
Fabre, E. F., Mouratille, D., Bonnemains, V., Palmiotti, G. P., & Causse, M. (2024). Making moral decisions with artificial agents as advisors. A fNIRS study. *Computers in Human Behavior: Artificial Humans, 2*(2), Article 100096. https://doi.org/10.1016/j.chbah.2024.100096
Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press. https://doi.org/10.1515/9781503620766
Gabriel, I., Keeling, G., Manzini, A., & Evans, J. (2025). We need a new ethics for a world of AI agents. *Nature, 644*(8075), 38–40. https://doi.org/10.1038/d41586-025-02454-5
Glickman, M., & Sharot, T. (2025). How human–AI feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour, 9*(2), 345–359. https://doi.org/10.1038/s41562-024-02077-2
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108. https://doi.org/10.1126/science.1062872
Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin, 135*(4), 555–588. https://doi.org/10.1037/a0015701
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*. https://doi.org/10.48550/arXiv.2504.15236
Hyland, K. (1998). *Hedging in scientific research articles*. John Benjamins Publishing Company. https://doi.org/10.1075/pbns.54
Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports, 13*(1), 4569. https://doi.org/10.1038/s41598-023-31341-0
Krügel, S., Ostermaier, A., & Uhl, M. (2026). Justification optional: ChatGPT's advice can still influence human judgments about moral dilemmas. *AI and Ethics, 6*(1), 120. https://doi.org/10.1007/s43681-026-01005-6
Lakoff, G. (1975). Hedges: A study in meaning criteria and the logic of fuzzy concepts. In D. Hockney, W. Harper, & B. Freed (Eds.), *Contemporary research in philosophical logic and linguistic semantics* (pp. 221–271). Springer. https://doi.org/10.1007/978-94-010-1756-5_9.
Leach, C. W., Ellemers, N., & Barreto, M. (2007). Group virtue: The importance of morality (vs. competence and sociability) in the positive evaluation of in-groups. *Journal of Personality and Social Psychology, 93*(2), 234–249. https://doi.org/10.1037/0022-3514.93.2.234
Lorch, R. F., Jr., Lorch, E. P., Ritchey, K., McGovern, L., & Coleman, D. (2001). Effects of headings on text summarization. *Contemporary Educational Psychology, 26*(2), 171–191. https://doi.org/10.1006/ceps.1999.1037
Lu, Z., Li, P., Wang, W., & Yin, M. (2022). The effects of AI-based credibility indicators on the detection and spread of misinformation under social influence. *Proceedings of the ACM on Human-Computer Interaction, 6*(CSCW2), 1–27. https://doi.org/10.1145/3555562
Luttrell, A., Petty, R. E., Briñol, P., & Wagner, B. C. (2016). Making it moral: Merely labeling an attitude as moral increases its strength. *Journal of Experimental Social Psychology, 65*, 82–93. https://doi.org/10.1016/j.jesp.2016.04.003
Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117–124). ACM. https://doi.org/10.1145/2696454.2696458.
Mazeika, M., Yin, X., Tamirisa, R., Lim, J., Lee, B. W., Ren, R., Phan, L., Mu, N., Khoja, A., Zhang, O., & Hendrycks, D. (2025). Utility engineering: Analyzing and controlling emergent value systems in AIs (arXiv:2502.08640). *arXiv*. https://doi.org/10.48550/arXiv.2502.08640
Metzger, M. J., & Flanagin, A. J. (Eds.). (2007). *Digital media, youth, and credibility*. MIT Press. https://mitpress.mit.edu/9780262562324/digital-media-youth-and-credibility/.
Moscovici, S. (1980). Toward a theory of conversion behavior. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 13, pp. 209–239). Academic Press. https://doi.org/10.1016/S0065-2601(08)60133-1.
Moussaïd, M., Brighton, H., & Gaissmaier, W. (2015). The amplification of risk in experimental diffusion chains. *Proceedings of the National Academy of Sciences, 112*(18), 5631–5636. https://doi.org/10.1073/pnas.1421883112
Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues, 56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153
Panicek, D. M., & Hricak, H. (2016). How sure are you, doctor? A standardized lexicon to describe the radiologist's level of certainty. *American Journal of Roentgenology, 207*(1), 2–3. https://doi.org/10.2214/AJR.15.15895
Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. University of Texas at Austin. https://doi.org/10.15781/T29G6Z
Petty, R. E., Briñol, P., & Tormala, Z. L. (2002). Thought confidence as a determinant of persuasion: The self-validation hypothesis. *Journal of Personality and Social Psychology, 82*(5), 722–741. https://doi.org/10.1037/0022-3514.82.5.722
Reynolds, C. J., Knighten, K. R., & Conway, P. (2019). Mirror, mirror, on the wall, who is deontological? Completing moral dilemmas in front of mirrors increases deontological but not utilitarian response tendencies. *Cognition, 192*, Article 103993. https://doi.org/10.1016/j.cognition.2019.06.005
Skitka, L. J., Bauman, C. W., & Sargis, E. G. (2005). Moral conviction: Another contributor to attitude strength or something more? *Journal of Personality and Social Psychology, 88*(6), 895–917. https://doi.org/10.1037/0022-3514.88.6.895
Skitka, L. J., Hanson, B. E., Morgan, G. S., & Wisneski, D. C. (2021). The psychology of moral conviction. *Annual Review of Psychology, 72*(1), 347–366. https://doi.org/10.1146/annurev-psych-063020-030612
Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI conference on human factors in computing systems*. https://doi.org/10.1145/3290605.3300768
Thomson, J. J. (1985). The trolley problem. *The Yale Law Journal, 94*(6), 1395–1415. https://doi.org/10.2307/796133
Tormala, Z. L., & Petty, R. E. (2002). What doesn't kill me makes me stronger: The effects of resisting persuasion on attitude certainty. *Journal of Personality and Social Psychology, 83*(6), 1298–1313. https://doi.org/10.1037/0022-3514.83.6.1298
Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*(2), 326–334. https://doi.org/10.1016/j.cognition.2012.10.005
Wood, W., Lundgren, S. R., Ouellette, J. A., Busceme, S., & Blackstone, T. (1994). Minority influence: A meta-analytic review of social influence processes. *Psychological Bulletin, 115*(3), 323–345. https://doi.org/10.1037/0033-2909.115.3.323