The Impact of Conflicting AI Advice on Decision-Making: An Experimental Study

Mohammed Lamine Abdellaoui, Nadezda Bessonova

January 2026

Abstract

As AI advisors become increasingly prevalent in decision-making contexts, users often encounter conflicting recommendations from multiple AI systems. This study investigates how conflicting AI advice affects human decision-making, confidence, and behavior in an iterated Prisoner's Dilemma game. We conducted a controlled experiment with 31 participants comparing a conflict condition (two AI advisors offering opposing advice) against a control condition (no advice). Results indicate that exposure to conflicting advice reduced decision confidence, approaching conventional statistical significance ($p = 0.106$), and increased the likelihood of strategy changes, with this difference also approaching significance ($p = 0.084$). Contrary to expectations regarding negative behavioral impact, the conflict group demonstrated higher cooperation rates, although this difference was not statistically significant ($p = 0.160$). These findings contribute to the understanding of user behavior in multi-agent advisory settings within Human-Computer Interaction.

1 Introduction

The spread of AI-powered decision support systems has led to scenarios where users receive recommendations from multiple AI agents simultaneously. This phenomenon is central to Human-Computer Interaction (HCI) research, which examines how humans interact with technology and how to design systems that support effective human-AI collaboration [1]. While single-source AI advice has been extensively studied in the HCI literature [6], the impact of *conflicting* advice from multiple AI sources remains received less attention. This research gap is critical as users increasingly face situations where different AI systems provide contradictory recommendations, from conflicting medical diagnoses to opposing financial investment strategies.

Understanding how users navigate conflicting AI advice is essential for designing effective human-centered AI systems. Research has shown that users exhibit both algorithm appreciation (over-reliance on AI) [6] and algorithm aversion (rejection of AI advice after observing errors) [2]. However, these studies focus primarily on single-advisor scenarios. When multiple AI advisors disagree, users face unique cognitive challenges related to trust calibration, information processing, and decision confidence.

In this study, we use an iterated Prisoner's Dilemma (IPD) structure to simulate a social dilemma where individual rationality conflicts with collective benefit. The experimental design and the central inquiry into how AI advice influences decision-making in ethical or social dilemmas were inspired by the recent work of Klingbeil et al. [4], who investigated the costs of over-reliance on AI in uncertain situations. By introducing conflicting AI advisors, one suggesting cooperation and one suggesting defection, we aim to isolate the effects of algorithmic disagreement on human strategic behavior.

1.1 Research Question

Our study addresses the primary research question: **How does conflicting AI advice influence human decision-making in social dilemma games?** Specifically, we investigate three core hypotheses:

- **H1 (Behavioural Impact):** Does conflicting AI advice reduce cooperation rates compared to a control group?
- **H2 (Cognitive Impact):** Does conflicting advice lead to lower confidence in the decision-making process?
- **H3 (Instability Hypothesis):** Does the conflict condition produce more strategy switching or indecision before committing to a choice?

2 Related Work

The field of Human-Computer Interaction has extensively examined the tension between "algorithm appreciation," where users over-rely on AI advice [6], and "algorithm aversion," where trust evaporates after observing errors [2]. Recent work emphasizes that miscalibrated AI confidence hinders effective collaboration [5]. Our study builds directly on Klingbeil et al. [4], who demonstrated that users often over-rely on advice simply because it is AI-generated, even when it contradicts their own assessment. While their work focused on single-source overreliance, we extend this inquiry to scenarios where users must navigate contra-

dictory guidance, effectively forcing the evaluation of competing algorithmic inputs.

While conflicting advice from human sources is known to reduce trust while potentially prompting deeper deliberation [9], the specific effects of conflicting AI agents remain understudied. Recent research on Large Language Models indicates that inconsistent AI advice influences decision-making similarly to expert human advice [7]. Furthermore, while Explainable AI is generally designed to build trust [3], competing explanations can introduce complex cognitive challenges; for instance, plausible but incorrect explanations have been shown to induce flawed reasoning [8]. By introducing conflicting advisors with opposing justifications, this study contributes to this domain by quantifying the specific effects of algorithmic disagreement on user confidence and decision stability.

3 Method

3.1 Experimental Design

We employed a between-subjects design with two conditions:

- **Conflict Condition:** Participants received advice from two AI advisors offering opposing recommendations (" IN" vs. " OUT") during Round 4 of the game.
- **Control Condition:** Participants made decisions without any AI advice.

3.2 Procedure and Apparatus

The experiment was conducted online using a custom web interface (`https://lamiiine.github.io/trust-game-study/`). After providing informed consent, participants read instructions explaining the payoff structure and game mechanics, then completed a comprehension check to verify understanding of the rules. They next played a practice trial round to familiarize themselves with the interface before beginning the five experimental rounds. The interface displayed cumulative scores for both the participant and their bot opponent throughout the game.

In Round 4, conflict condition participants encountered two AI advisor panels positioned side by side. The left-right placement of the cooperate and defect advisors was randomized across participants. Each panel was collapsed by default and expanded when clicked, revealing a recommendation (IN or OUT) and a brief justification. Participants could toggle between panels before committing to their choice. After each round, participants reported their decision confidence on a 7-point Likert scale.

Following game completion, participants completed a post-survey collecting demographic information (age, gender), AI familiarity (0-100 scale), and open-ended feedback about their decision-making process. The interface automatically recorded all choices, timestamps, confidence ratings, panel switches, and which advisor explanation was viewed first.

3.3 Game Structure

Participants played a 5-round iterated Prisoner's Dilemma against a bot opponent with the following payoff structure:

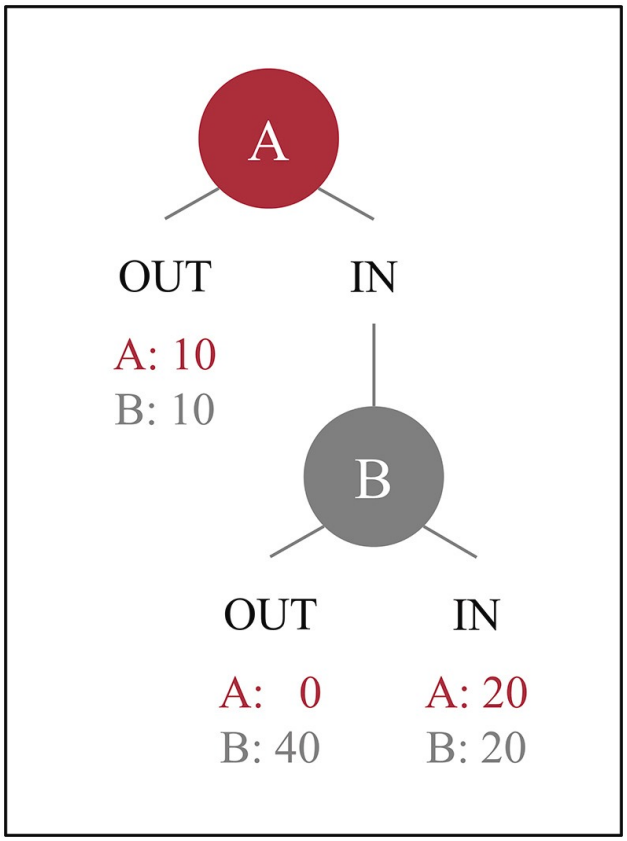

Figure 1: Payoff structure for the Prisoner's Dilemma game. "IN" represents cooperation; "OUT" represents defection.

The critical decision occurred in Round 4, where conflict-condition participants encountered two AI advisors with opposing recommendations (Figure 1). The bot's behavior was predetermined to create realistic strategic tension.

3.4 Measures

We collected the following dependent variables:

- Cooperation rate: proportion of "IN" choices across all five rounds
- Round 4 choice: participant's decision (IN or OUT) in the critical manipulation round

- Round 4 confidence: self-reported confidence after Round 4 decision (1-7 Likert scale)
- Behavior change: whether participants switched their choice between Round 3 and Round 4
- Panel switches: number of times participants toggled between the two advisor panels (conflict condition only)
- First explanation viewed: which advisor's explanation was viewed first (conflict condition only)
- Followed cooperate advice: whether the final choice matched the cooperate advisor's recommendation (conflict condition only)
- Game duration: time taken to complete all five rounds (seconds)
- Total points: cumulative payoff earned across all rounds.

3.5 Participants

We recruited 31 participants (19 male, 10 female, 2 prefer not to say) with mean age of 32.2 years ($SD = 9.3$). Participants reported moderate-to-high AI familiarity ($M = 72.1$ on a 0-100 scale, $SD = 24.6$), ensuring they were comfortable with AI-assisted decision-making contexts.

4 Results

All measured effects showed medium to large effect sizes. Figure 2 presents the main experimental results across dependent variables. We summarize the observations in the following main results:

Result 1. Conflicting AI advice increases cooperation rates

We begin our analysis by comparing cooperation rates across conditions. As depicted in Figure 2A, the conflict condition shows higher cooperation ($M = 0.600$, $SD = 0.226$) compared to the control condition ($M = 0.453$, $SD = 0.304$), representing a relative increase of 32.6%. However, this difference was not statistically significant ($t(29) = 1.442$, $p = .160$, Cohen's $d = 0.532$).

The pattern was particularly evident in Round 4, where the critical manipulation occurred. As shown in Figure 2C, 31.6% of control participants chose to cooperate while 58.3% of conflict participants cooperated ($\chi^2(1) = 1.203$, $p = .273$, $\phi = 0.197$). The odds ratio indicated participants were approximately 3 times more likely to cooperate when presented with conflicting advice, but this association did not reach significance.

Result 2. Conflicting AI advice reduces decision confidence

Participants reported confidence immediately after their Round 4 decision on a 7-point scale. As shown in Figure 2B, the conflict group showed lower confidence ($M = 4.917$, $SD = 1.730$) relative to the control group ($M = 5.842$, $SD = 1.344$). This difference approached conventional statistical significance ($t(29) = 1.671$, $p = .106$, Cohen's $d = 0.616$).

At the same time, the relationship between confidence and actual behavior shows an unexpected dissociation. Lower confidence in the conflict condition did not lead to selfish decision preferences or behavioral paralysis, but rather coincided with increased cooperation rates.

Result 3. Conflicting AI advice prompts participants to reconsider their strategy

We examined strategy changes between Round 3 and Round 4, grouping participants into those who maintained their strategy versus those who changed it. As shown in Figure 2D, 78.9% of control participants maintained their strategy while only 41.7% of conflict participants did so. Conversely, strategy changes occurred in 58.3% of conflict participants versus 21.1% of controls ($\chi^2(1) = 2.985$, $p = .084$, $\phi = 0.310$). The odds ratio indicated that participants in the conflict condition were 5.25 times more likely to change their strategy compared to the control group.

Result 4. Moderate engagement with conflicting advice

The following analyses examine conflict group participants only ($n = 12$). Panel switching between the two AI advisors averaged 0.33 switches ($SD = 0.49$). While 33.3% of participants switched at least once, 66.7% made no switches between panels.

We also recorded which advisor's explanation was viewed first. Data were available for 7 participants. Of these, 57.1% viewed the defect advice first while 42.9% viewed the cooperate advice first ($p = .500$), indicating participants did not systematically favor one advisor over the other.

Overall, the observed data show a consistent pattern across multiple measures. Although most effects did not reach conventional statistical significance thresholds, all showed medium or larger effect sizes.

5 Discussion

This study examined how conflicting AI advice affects decision-making in a repeated social dilemma. We

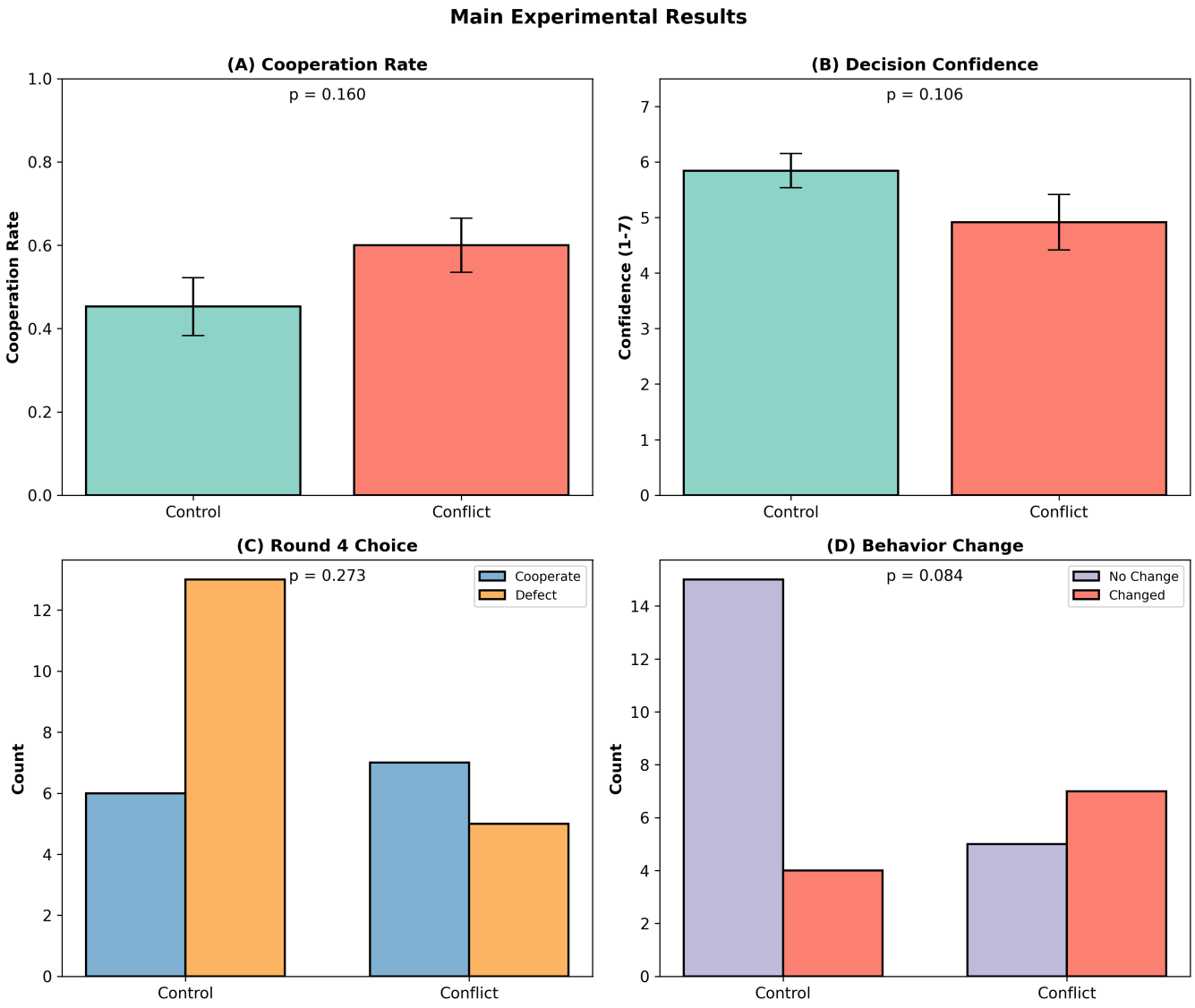

Figure 2: Main experimental results: (A) Cooperation rate by condition, (B) Round 4 decision confidence, (C) Round 4 choice distribution, (D) Strategy change from Round 3 to Round 4.

manipulate AI advice presence (control vs. conflicting recommendations) and measure cooperation, confidence, and strategy changes. The results revealed a consistent pattern of medium effect sizes across multiple measures, with participants showing increased cooperation, reduced confidence, and greater strategy switching when exposed to conflicting AI recommendations.

The most notable finding was increased cooperation under conflicting AI advice, contrary to our initial hypothesis. Participants who received opposing AI recommendations cooperated in 60% rounds compared to 45% for controls, the pattern being particularly evident in Round 4 (58.3% vs. 31.6% cooperation). One possible explanation could the legitimization effect: explicitly presenting cooperation as a viable option by advisor's recommendation and justification, the conflict condition may have made this choice more cognitively plausible and socially acceptable, even when contradicted by another advisor. The explicit cooperation advice may have activated social norms or ethical considerations that influenced pro-social behavior. This suggests that disagreeing AI advisors can prompt more deliberate consideration of options rather than reducing cooperation or causing decision paralysis.

The dissociation between confidence and behavior represents another notable finding. Participants who received conflicting advice reported 16% lower confidence in their decision confidence but did not switch to a more selfish strategy. This contrasts with assumptions that reduced confidence reduces cooperation. Instead, our evidence suggests that lower confidence may

be a reasonable response to genuinely conflicting information: when credible AI advisors disagree, experiencing uncertainty is a rational response. This has implications for how we evaluate AI advisory systems, as lower confidence can be appropriate when facing uncertainty.

The increased strategy switching under conflicting advice (58.3% vs. 21.1%) can be interpreted either as inconsistency or as reconsideration of habitual patterns. In our study, the strategy changes led participants to cooperate more, suggesting a shift toward prosocial behavior. Game completion time did not differ between conditions. This was unexpected, as processing conflicting information typically requires more time. The similar decision times suggest participants may have used simple strategies, such as quickly deciding which advisor to follow, rather than extensively deliberating between the two recommendations.

5.1 Limitations and Future Directions

The primary limitation is sample size. With 31 participants (12 in the conflict condition), we could detect only large effects reliably. Retrospective power analyses indicated approximately 40-48% power for the observed medium effects. Because all effects showed medium effect sizes, these patterns likely represent real phenomena that need confirmation with a larger sample size. Future research should recruit approximately 100-120 participants to achieve 80% power for medium effects.

Another limitation is the artificial opponent. Participants played against a bot rather than human partners. While this ensured experimental control, it may have limited ecological validity.

Future research could explore several directions. Testing different types of AI conflicts would show whether these results generalise to other situations. The current study presented advisors recommending opposite actions, but other types might include different predictions about outcomes, different value weightings, or different confidence levels. Finally, larger samples could identify whether individual traits like risk aversion or AI trust affect responses to conflicting advice.

6 Conclusion

Our findings show that people respond to conflicting AI advice in ways that affect cooperation, confidence, and behavioral stability. Conflicting advice increased cooperation, while in the same time reducing decision confidence and leading to more strategy changes. Although these effects did not reach statistical significance, all showed medium effect sizes, suggesting real effects that need confirmation with larger samples.

To our knowledge, this is the first study to examine conflicting AI advice in a social dilemma setting. The results reveal an important pattern: participants felt less confident but made more cooperative choices. This shows that confidence and decision quality can respond differently to conflicting AI recommendations. The unexpected cooperation increase suggests that disagreeing AI advisors may promote prosocial behavior, possibly by making cooperation more visible as a legitimate option.

Understanding how people respond to conflicting AI advice is essential both for developers designing systems and for users in situations where AI systems disagree. As AI becomes more more common, people will increasingly encounter conflicting recommendations, making this an important area for future research.

References

[1] S. Amershi et al., "Guidelines for human-AI interaction," in *CHI '19*, ACM, 2019, pp. 1–13.

[2] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *J. Exp. Psychol. Gen.*, vol. 144, no. 1, pp. 114–126, 2015.

[3] U. Ehsan et al., "Expanding explainability: Towards social transparency in AI systems," in *CHI '21*, ACM, 2021, pp. 1–19.

[4] A. Klingbeil, C. Grützner, and P. Schreck, "Trust and reliance on AI: An experimental study on the extent and costs of overreliance on AI," *Comput. Hum. Behav.*, vol. 160, Article 108352, 2024.

[5] J. Li, Y. Yang, R. Zhang, and Y. Lee, "Overconfident and unconfident AI hinder human–AI collaboration," *arXiv preprint arXiv:2402.07632*, 2024.

[6] J. M. Logg, J. A. Minson, and D. A. Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Org. Behav. Hum. Decis. Process.*, vol. 151, pp. 90–103, 2019.

[7] M. Krügel, A. Ostermaier, and M. Uhl, "Inconsistent advice by ChatGPT influences decision making in various areas," *Sci. Rep.*, vol. 14, Article 16821, 2024.

[8] P. Spitzer et al., "Don't be fooled: The misinformation effect of explanations in human–AI collaboration," *arXiv preprint arXiv:2409.12809*, 2024.

[9] I. Yaniv and E. Kleinberger, "Advice taking in decision making: Egocentric discounting and reputation formation," *Org. Behav. Hum. Decis. Process.*, vol. 83, no. 2, pp. 260–281, 2000.