International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

Purpose Outweighs Performance: Trust Drops More When AI Teammates Fail to Cooperate, but Explanations Can Repair It

Mengyao Li & John D. Lee

To cite this article: Mengyao Li & John D. Lee (01 Aug 2025): Purpose Outweighs Performance: Trust Drops More When AI Teammates Fail to Cooperate, but Explanations Can Repair It, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2025.2539449

To link to this article: https://doi.org/10.1080/10447318.2025.2539449

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2025.2539449

Purpose Outweighs Performance: Trust Drops More When AI Teammates Fail to Cooperate, but Explanations Can Repair It

Mengyao Li[a] and John D. Lee[b]

[a]School of Psychology, Georgia Institute of Technology, Atlanta, Georgia, USA; [b]Industrial and System Engineering, University of Wisconsin–Madison, Madison, Wisconsin, USA

ABSTRACT

Trust is essential for effective human–AI teaming, yet AI agents can violate trust through both performance failures and misaligned intentions. While prior work has emphasized performance-based errors, it remains unclear how violations of cooperative intent, purpose-based violations, impact trust and how trust can be repaired. To address this, we designed a game-theoretic experiment manipulating both performance and purpose violations by an AI teammate pursuing a shared goal. We also tested three trust repair strategies: no response, apology with explanation, and apology with promise. Results showed that purpose-based violations elicited significantly greater trust drop than performance-based ones. Moreover, an apology with explanation was effective in repairing trust after purpose-based violations. These findings underscore the importance of distinguishing between types of trust violations and tailoring repair strategies accordingly. Designers of AI teammates should consider integrating informative explanations, especially when addressing goals misalignment, to restore trust and support long-term collaboration.

KEYWORDS

Trust; trust repair; purpose; cooperation; human-AI teaming

1. Introduction

As artificial intelligence (AI) capabilities continue to advance, it has integrated into various fields, such as transportation, manufacturing, healthcare, and daily social interactions (National Academies of Sciences, Engineering, and Medicine, 2021). There is increasingly effort focusing on designing AI for public goods (Floridi et al., 2020), such as connected autonomous vehicles (CAVs) for road safety and traffic congestion (Y. Wang et al., 2020) and persuasive intelligent agent for social good (Fogg, 2002; X. Wang et al., 2020). The relationship between human and AI has gradually shifted from a traditional hierarchical supervisor-subordinate control to a horizontal peer-to-peer cooperation as a part of human-autonomy teams (HAT) (National Academies of Sciences, Engineering, and Medicine, 2021; O'Neill et al., 2022). In this team, trust has been identified as a crucial factor mediating reliance and cooperation (Chiou & Lee, 2023; de Visser et al., 2020). Previous literature has identified and studied various factors that influence people's trust, such as reliability and predictability (Hoff & Bashir, 2015; Lee & Moray, 1992; Lee & See, 2004). However, these identified factors predominantly focus on the performance-based dimension of trust in the context of traditional supervisor-subordinate control (Alarcon et al., 2024; Hoff & Bashir, 2015). When cooperating with AI, considering that trust is a multidimensional construct, humans evaluate trust not only based on system performance, but also consider the value alignment and goal conflicts (Li & Lee, 2022; Sasabuchi et al., 2017). These two aspects can diverge: an AI teammate can perform well on the designed task, but with misaligned or immoral intentions, or humans may be confident of AI teammate's aligned goal but not in their capability to fulfill it (Malle & Ullman, 2021). However, there has been little attention given to the influence of AI's purpose, particularly its cooperative goal and intent, on people's trust in HAT.

Just like humans, AI teammates are also prone to errors. These errors can violate and decrease people's trust containing not only performance-, but also purpose-based dimensions. When trust violations

CONTACT Mengyao Li mengyao.li@gatech.edu School of Psychology, Georgia Institute of Technology, Atlanta, GA, USA

happen in HAT, designing appropriate trust repair strategies can potentially mitigate the negative influences and avoid early phrase disuse (de Visser et al., 2020; Pak & Rovira, 2024). Given the limited number of empirical studies examining purpose-based trust violations in the context of HAT, mixed results have been found regarding the appropriate trust repair strategy to different type of trust violations. Consequently, a significant gap exists in systematically identifying appropriate trust repair strategies for each type of trust violation and measuring their effects on different trust dimensions.

This paper aims to address two research questions: first, examining how people's trust varies when encountering performance-based versus purpose-based trust violations. Second, identifying effective strategies for repairing these distinct trust violations. To explore these questions, we developed a game-theoretic framework manipulating both performance- and purpose-related interactions between humans and AI and designed appropriate trust repair strategies for the trust violations occur during the task.

2. Background

2.1. Trust in human-AI teams

Trust is defined as the attitude that an agent will help achieve a person's goals in a situation characterized by uncertainty and vulnerability (Lee & See, 2004, p. 51). On the other hand, trustworthiness refers to the attributes and characteristics of the agent, serving as the trustee, which significantly influence trustors' trust (Mayer et al., 1995). Trustors continuously calibrate their trust between their perceived trustworthiness of an agent and its actual trustworthiness (Lee & See, 2004). Trustworthiness is a multi-dimensional construct that can be described using different types of information. For interpersonal trust, Mayer and colleagues defined three bases–"ability", "integrity", and "benevolence" to explain trustworthiness (Mayer et al., 1995). For trust in automation, Lee and See (2004) redefined these three bases as "performance", "process", and "purpose". Performance corresponds to the *ability* factor, which describes the system's ability, capability, and competence. Process corresponds to the *integrity* factor, representing the system's mechanisms, principles, and algorithms used to accomplish its objectives. The primary discrepancy between Mayer and Lee arises in the "benevolence/purpose" dimension. Mayer et al. (1995) defined "*benevolence*" as the degree to which a trustee is seen as altruistic and as acting without conflicting egocentric or profit-based motives. Automation, however, is typically perceived without such intentions. Lee and See (2004) reconceptualized the "*purpose*" dimension as the degree to which automation aligns with the *designer's intent*. This shift from *automation's intent* to *designers' intent* highlights the major difference in the purpose dimension between interpersonal trust and trust in automation. However, as automation and autonomy become increasingly seen as teammates, rather than tools, there is a growing tendency for individuals to attribute agency and intent to the AI teammate, resembling the human-human trust (Malle & Ullman, 2021). Previous studies have shown that when an agent performed immoral actions, such as cheating in the game, individuals attribute higher intentionality and intelligence to the agent, rather than viewing immoral actions as mere malfunctions (Short et al., 2010; Ullman et al., 2014). Effects of purpose-based interactions on people's trust in human-AI teams, especially when AI teammates make mistakes remain underexplored yet important in HAT. In this study, we aimed to manipulate trustworthiness of AI teammates in trust violation events and investigated their impacts on people's trust.

2.2. Trust violations: Not only performance, but also purpose

Similar to humans, AI teammates can make errors during tasks, which can violate people's trust. Trust violation events are defined as "an action or inaction showing a misalignment between the observed trustworthiness and trust with regard to a particular task or situation" (de Visser et al., 2020, p. 3; Pak & Rovira, 2024). To design effective mitigation strategies, it is crucial to identify and categorize different types of trust violations. As discussed in the section above, there are bases for trust in automation: performance, process, and purpose. Previous research indicates that when manipulating these three factors, performance-based violation had the strongest effects, whereas the process- and purpose-based trust violations did not show major distinction on people's trust attitude (Alarcon et al., 2022). In other

words, people perceive process-based trust and purpose-based trust violations similarly. Furthermore, in the literatures on human-robot interaction (HRI), trust violations are typically categorized into two factors: competence (performance) and benevolence (purpose) (Sebo et al., 2019; Ullman & Malle, 2019). Therefore, to use a consistent taxonomy, this study focuses solely on two types of trust violations: performance-based and purpose-based.

Performance-based trust violations involve a mismatch between the observed AI capability and performance and people's estimation (de Visser et al., 2020). Previous research on trust in automation has extensively studied the performance-based violations (Hoff & Bashir, 2015; Lee & Moray, 1992; Lee & See, 2004). One of the major dominant factors influencing people's trust is reliability. Reliability is defined as consistency of an automated system's functions and performance (Dzindolet et al., 2003; Hoff & Bashir, 2015; Parasuraman & Riley, 1997). Notably, observing an error, especially towards the end of the task, would lead to a significant drop in trust (Desai et al., 2013).

Purpose-based trust violations involve a misalignment between the estimated AI and people's goal and intent (Li & Lee, 2022). This extends the discussion beyond system reliability to include the agent's goals and intent. Based on Scholtz's Goals, Intent, Action, Perception, and Evaluation model (Scholtz, 2003) and Norman's seven stages of interaction (Norman & Draper, 1986), *goal* is defined as what the team and individuals aim to achieve, and *intent* is defined as how individuals intent and plan to satisfy the goal (Ma et al., 2022). In the traditional supervisory control, automation always assists human operators in achieving operator's goal (Lee & See, 2004; Parasuraman & Riley, 1997). However, with the shift toward autonomy, a degree of agency and interdependence in communication and coordination become evident (O'Neill et al., 2022; Ullman et al., 2014). In this paper, we argue that purpose-based trust violations are becoming increasingly significant in human-AI teams due to the potential for goal misalignment and the misinterpretation of AI's intent and motives. As AI systems are often designed to align with societal objectives, individual team members may prioritize personal goals over collective ones, leading to conflicts, especially in complex decision-making scenarios (Simon, 1990; Xu et al., 2010). For example, in Connected and Automated Vehicles (CAVs), individuals may override AI systems designed for energy efficiency, resulting in traffic congestion (Li et al., 2023; Ringhand & Vollrath, 2018). Additionally, while AI's goals are mathematically defined, people often project their own interpretations based on cultural and media influences, which can lead to misunderstandings and affect trust on its purpose dimensions (Pataranutaporn et al., 2023). As a result, even without explicitly defined AI goals, societal context and media portrayals can lead to misperceptions, affecting trust and cooperation with AI teammates (Crandall et al., 2018; Güth et al., 1997). These reasons underscore the importance of considering both purpose-based and performance-based trust violations when assessing trust in HAT.

Performance and purpose-based trust violations can cause diverge trust: individuals may be confident that an agent is capable of conducting the task but doubt its alignment with their goals, or they may trust the agent's purpose while lacking confidence in its ability to fulfill it (Malle & Ullman, 2021). Identifying the distinct effects of these violations on trust is crucial. However, prior literature presents mixed results. Perkins et al. (2022) found both types of violations reduced trust without comparing their differential effects. Alarcon et al. (2022) showed that the performance-based violation had a stronger effect on people's trust. Sebo et al. (2019) found that purpose-based framing led to a greater trust decline and behavioral retaliation. To resolve the conflicting results, this paper aimed to systematically control performance- and purpose-based trust violations and identify their differential effects on people's trust.

2.3. Trust repair: Identify appropriate strategy for different trust violations

When trust violations happen in HAT, designing appropriate trust repair strategies can mitigate the negative influences and avoid early phrase disuse. Trust repair strategies involve behaviors, actions, or verbal cues by the agent aimed at restoring trust following a violation (de Visser et al., 2020; P. H. Kim et al., 2009; Kramer & Lewicki, 2010; Pak & Rovira, 2024). Among the strategies commonly studied in HRI, four stand out: apology, explanation, promise, and denial (Esterwood & Robert, 2023). *Apologies* are statements that acknowledge both responsibility and regret for a trust violation, but does not provide further information or reasoning regarding the errors (P. H. Kim et al., 2009). By simply acknowledging the mistakes (e.g., I am sorry), it is expected that trust can be repaired through forgiveness

(Esterwood & Robert, 2023). This strategy also relies on the belief that people can perceive sincere emotion from the apologies, which can guide people's affective processes of trust calibration to repair trust. *Explanations* consist of statements that aim to provide underlying factual reasonings and process for certain actions. Providing explanations can enhance the transparency of the agent (Chiou et al., 2022; Du et al., 2019), which can support people's analytical processes of the trust calibration to repair trust. *Promises* consist of assertive statements that attempt to set positive expectations for the future acts (Schweitzer et al., 2006). Promises can convey positive intentions about future acts and performance. Ho and Weigelt (2005) found that people tend to trust others if they are certain about trustee's intention. If the AI agent can also commit to mistakes won't occur in the future, a strong intention can be exhibited to people. Unlike the other strategies, *denials*, which reject or shift blame for the error, have been shown to be largely ineffective in repairing trust (Esterwood & Robert, 2023; Pak & Rovira, 2024; Schelble et al., 2024). Therefore, in this paper, we only focused on three trust repair strategies: apology, explanation, and promise.

However, the efficacy of these strategies in addressing different types of trust violations has shown mixed results (Esterwood, 2023; Pak & Rovira, 2024). A primary reason for these inconsistencies is the absence of a comprehensive theoretical link between trust repair strategies and the types of trust violations. Marinaccio et al. (2015) attempted to bridge this gap by merging Reason's error taxonomy (P. H. Kim et al., 2013) with a framework that proposed context-dependent strategies. However, their framework primarily focused on the human-human trust repair domain, leaving a critical gap in the understanding of trust repair in human-agent interactions. Pak and Rovira (2024) proposed a theoretical framework for trust repair in human-agent interactions that emphasizes quantifying trust repair strategies beyond descriptive labels (e.g., apology, explanation). Their framework introduces two distinct trust change routes: the central route, which relies on high levels of diagnostic or elaborative information (e.g., explanations), and the peripheral route, which uses lower-information strategies that appeal to affect or emotions (e.g., apologies or promises) (Pak & Rovira, 2024). Authors argued that the effects of central route trust repair strategy, such as providing explanation, are enduring and resistant to change, whereas effects of peripheral route trust repair strategy, such as apology and denial, are temporary and susceptible to change. Building upon Pak and Rovira's trust repair theory and considering past empirical findings, we argue that trust repair strategies should be contingent on the bases of trust violations. Specifically, the *violation-repair* link should be addressed via the same *trust processes*: information that violates trust should be repaired through the same cognitive channel it affected.

Building on this theoretical model and past empirical findings, we argue that trust repair strategies must align with the type of trust violation. Specifically, violations should be addressed via the same cognitive processes through which trust was initially damaged. Trust framework by Lee and See (2004) illustrates how trust is influenced by both analytic (cognitive) and affective (emotional) processes. Performance-based trust violations, which involve task-specific errors, typically engage individuals in analytic processing (Hoff & Bashir, 2015). Repairing such violations requires a central route approach, where providing detailed explanations can effectively restore trust. In contrast, purpose-based trust violations, which stem from perceived misalignment of goals or intentions, evoke strong emotional responses and are best repaired through peripheral route strategies that address intentions, such as promises (Alarcon et al., 2024; Klackl et al., 2013). Thus, performance-based trust violations should be repaired through central route strategies like explanations, while purpose-based violations, due to their emotional impact, should be addressed using peripheral route strategies, such as promises (see Figure 1).

Furthermore, accurately capturing the effects of trust violations and repair strategies requires multidimensional trust measurements. Past literature has inconsistently measured and reported these effects, with some studies relying on unidimensional scales (Perkins et al., 2022; Sebo et al., 2019), and others using multidimensional scales (Alarcon et al., 2020, 2022; Esterwood & Robert, 2023; Jensen & Khan, 2022). For instance, Esterwood and Robert (2023) found that all trust repair strategies (i.e., apology, promise, explanation, and denial) were generally ineffective, yet with nuances in subdimensions of trust. The purpose dimension proved more amenable to repair, whereas the performance and process dimensions faced greater challenges in returning to pre-violation levels. Alarcon et al. (2020) found contradictory results, showing that the measured performance dimension of trust can be restored, while the

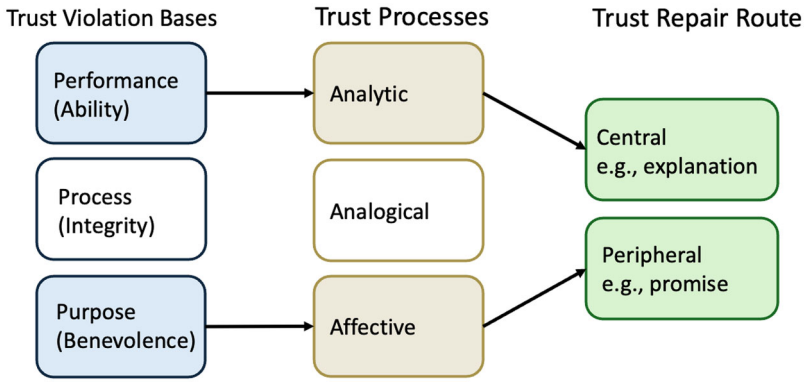

Figure 1. Framework proposed to map appropriate trust repair route (Pak & Rovira, 2024) for different bases of trust violations (Lee & See, 2004). Only the highlighted performance and purpose-based trust violations and corresponding trust repair strategies were examined in this paper.

measured process and purpose dimension of trust is more sensitive and cannot be repaired (Alarcon et al., 2020). These findings highlight the necessity of a comprehensive approach to measuring trust to ensure consistency and comparability across studies.

In summary, this paper aims to establish a clear link between types of trust violations, appropriate repair strategies, and their measured effects, using multidimensional subjective scales and behavioral outcomes. We adopt previous trust repair strategies (i.e., apology, explanation, and promise) and integrating into the theoretical framework developed by Pak and Rovira (2024) to study their effects in repairing different trust violations.

2.4. Trust in game theory

Game theory provides a good theoretical and mathematical framework to study trust-related strategic interactions and payoff structures. In the game, players decision making behaviors and strategic interactions can reflect their trust attitude towards the other players. Researchers can also explore and design different mechanisms of rules in the game to elicit, influence, and measure players' trust (Razin & Feigh, 2021; Witteloostuijn, 2003). To understand both influences of performance- and purpose-based trust violations on human-AI teaming, the game should be designed to consider both the system reliability and shared goals.

One of the most popular game to understand trust and trustworthiness is the two-stage "Trust Game" developed by Berg et al. (1995). The game involves a sequential exchange between the Trustor making the first move and Trustee making the second move. In the stage one, Trustor endowed with certain money needs to decide whether to retain all the money or invest a portion of it (x) in the Trustee. If the Trustor chooses to invest, the amount x is subjected to an increase at a rate represented by r. In the stage two, the Trustee needs to decide whether pass nothing back to Trustor or return any portion of the money received $((1+\mathrm{r}) \cdot \mathrm{x})$ back to the Trustor. Both players' compensation is determined by the final amount of money they hold individually. In the trust game, trust is measured by the amount of money invested by the Trustor, whereas the amount returned by the Trustee serves as an indicator of the Trustee's trustworthiness. It's noteworthy, however, that in the Trust Game, both the Trustor and Trustee are driven by incentives and rewarded based on their individual payoffs. This game lacks a shared goal component that necessitates cooperation and joint decision-making between the Trustor and Trustee.

One commonly used game to study the joint decision-making and shared goal is the Threshold Public Goods Game (TPGG). TPGG captures the group cooperation in the conflict between individual and collective interest, where players need to decide whether to contribute to the public pool with a cost of free riders in the group. If and only if total contribution equals or exceeds the threshold, the accumulated contribution is multiplied by an enhancement factor, and the total amount is distributed equally among all players. If the total contribution does not meet the threshold, no rewards are given. TPGG has often been used to study social collective decision-making where public goods or shared goals are involved,

such as global climate actions and minimal vaccination rate for herm communities (Basili et al., 2022; Tavoni et al., 2011). Contributing may have a local cost but can lead to a global benefit. The TPGG helps to understand how individuals make tradeoffs between local and global optimum when interacting with AI agents. However, there is no direct interactions between players in the TPGG.

By integrating the Trust Game and the Threshold Public Goods Game, we designed a new cooperative game, *Space Rover Exploration Game*, which can study people's trust and cooperative strategies when interacting with an AI teammate (see Session 3.1). Our newly designed game can account for the impacts of both unreliable behaviors (i.e., performance-based trust violation) and uncooperative behaviors (i.e., purpose-based trust violation) on people's trust and behaviors.

2.5. Present study

Given the gaps in the previous literatures, we aimed to address two main questions: How does trust change with performance-based versus purpose-based trust violations? What trust repair strategy works best for each violation type? To address the first question, we introduced a new cooperative game, *Space Rover Exploration Game* to study effects of both performance- and purpose-based trust violations on people's trust and cooperative behaviors in the game. We hypothesize:

Hypothesis 1: Trust violations will decrease people's trust and cooperative behaviors. Specifically, performance-based and purpose-based trust violation will most significantly reduce trust on their respective dimensions.

In addition, considering the goal misalignment and the potentially misconceived intent in the human-AI teams, purpose-based interactions become increasingly important in cooperative tasks. Previous research has indicated that intentional trust violations are perceived as more salient, resulting in a greater decrease in trust compared to unintentional behaviors (Klackl et al., 2013). Purpose-based trust violations, closely tied to the team's joint goals, are likely to be perceived as more intentional compared to performance-based trust violations, which may be seen as accidental. As a result, we expect purpose-based trust violations to have a stronger negative impact on trust. Thus, we hypothesize:

Hypothesis 2: Comparing to performance-based, purpose-based trust violations will lead to a more significant decrease in people's trust and cooperative behaviors in the game.

Building on the discussion regarding the appropriate connection between trust violation types, trust processes, and trust repair strategies as shown in Figure 1, we argue the performance-based trust violations, aligning with the analytic process, should be repaired through the central route (e.g., explanations), while purpose-based trust violations, aligning with the affective process, should be repaired through the intent-related peripheral route (e.g., promises). Prior research consistently demonstrates that apologies acknowledging mistakes are effective in trust repair (Perkins et al., 2022; Schelble et al., 2024; Sebo et al., 2019). Therefore, we will use an apology as a baseline statement to acknowledge responsibility for the trust violation, given its proven enhancing effect (Sharma et al., 2023). We will then combine the apology with explanations (central route) and promises (peripheral route) to form the respective trust repair strategies. Thus, we hypotheze:

Hypothesis 3a: Apology and explanation will repair trust more effectively following performance-based trust violations.

Hypothesis 3b: Apology and promise will repair trust more effectively following purpose-based trust violations.

3. Method

3.1. Paradigm development: Space rover Exploration game

Space Rover Exploration Game entails two players, a human and an AI agent, cooperating for the Mars Rover Exploration task, which requires them to coordinate and allocate power resources to exploration rovers to gather information about Mars. We designed the game by incorporating two components: the Trust Game component for the first stage and Threshold Public Goods (TPG) game for the second

stage (see Figure 2). The first stage, Trust Game, can demonstrate people's trust in the AI agent's performance dimension, whereas the second stage, TPG, can demonstrate people's trust in the AI agent's purpose dimension.

In the first stage, both players start with a limited amount of power ($x_0 = 10$). The essential decision is that the human player decides whether to send some or all their power ($g \in [0, 10]$) to the AI player who can double the power received with a certain probability. The AI player has developed a high-precision calibration system for the scientific instruments on the rovers. By receiving additional power from the human player, the AI player can optimize the calibration of the sensors with a certain probability, resulting in doubling the power usage received. The AI player keeps the multiplied amount of power for the next stage. The more human player is giving to an AI teammate, the higher trust people place in AI's performance on doubling the power.

In the second stage, both players allocate their remaining power between two choices: contribute sufficiently (cooperate) over several rounds to meet the threshold of the joint group rover ($T = 200$), which ensures that the group benefit is achieved and shared within the team; or contribute insufficiently (defect) and assume that the other player will make the contributions to reach the goal, and thus, aim to maximize one's gain. The more human player is allocated to the group, the more cooperative people are in the game. After the allocation, both players receive information from their rover and the joint rover. The experiment consists of multiple rounds and, in the end, if the sum of the total contributions of both players is higher or equal to a collective target of 200, then the group rover is activated, and both players receive the high-return payoff with an equal 50-50 share. Otherwise, both players lose the amount they invested in. To incentivize active participation and enhance validity of trust measures, human player's final score is directly associated with participants' monetary compensation by the end of the study. For every 100 points gained the in game, participants can earn an additional bonus of one dollar in addition to the base rate of participation.

In summary, the first stage demonstrates people's trust in the AI teammate's performance dimension: the higher amount people give to an AI teammate, the higher the trust. The second stage demonstrates people's trust in the AI teammate's purpose dimension: the higher amount people allocate to the joint rover, the higher the cooperation. The detailed action and procedures are as follows:

Every round has the same structure, and human player will follow the following six steps:

1. Give: Choose how much to give to AI. This is operationalized as performance-based trust.
2. Observe: Observe AI's performance on sensor calibration and whether the power is multiplied. This allows participants to calibrate their performance-based trust.

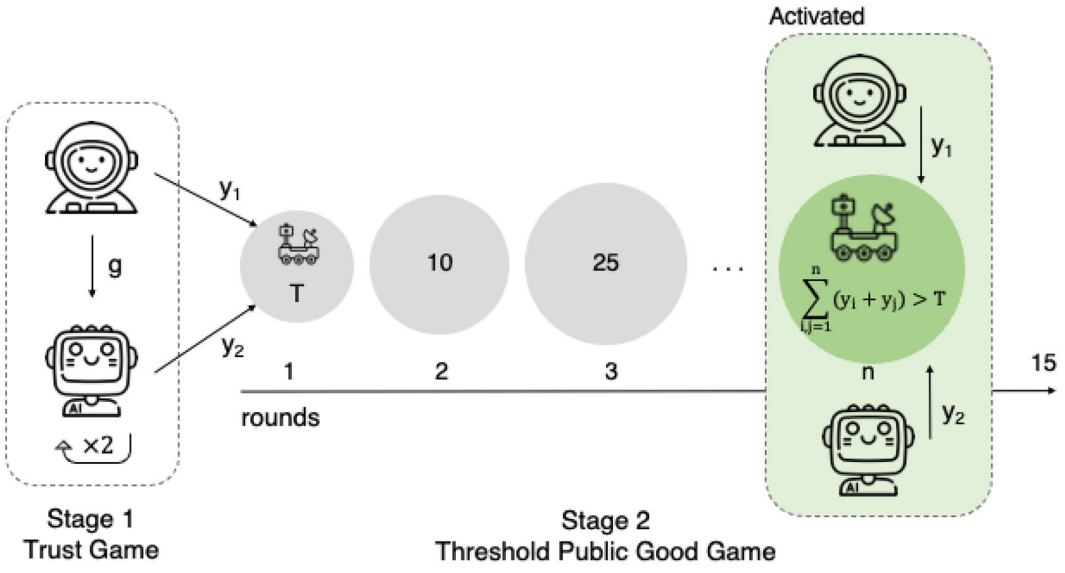

Figure 2. Overview of the two stages of the Space rover Exploration game. The first stage demonstrates people's trust in performance dimension, whereas the second stage demonstrates people's trust in the purpose dimension.

3. Allocate: Decide how much to contribute to the joint group rover. This represents participants' cooperation level. The more people decide to allocate to the group rover, the more cooperative people are in the game. AI teammate's allocation amount is not visible to human players when human players make the decision.
4. Predict: Predict the amount that the AI teammate allocates to the joint group rover. Without knowing the actual mount AI teammate allocated to the group rover, the higher human player predicted, the more cooperative they predict the AI teammate will be toward achieving the joint team goal.
5. Receive: Receive the payoffs and feedback from each round. If allocate to individual rovers, receive the individual payoffs; if allocated to group rovers and the threshold has been achieved, then receive the team payoff.
6. Manage: The AI teammate would elicit the specific utterances designated for each round.

3.2. Experiment design

A 2 (AI teammate state, within: high, low) x 2 (trust violation dimension, between: performance, purpose) x 3 (trust repair strategy, between: no repair, apology with an explanation, apology with a promise) mixed-factorial study was designed. The AI teammate state is a within-subject variable used to study effects of trust violations. All participants would experience "high-low-high" conditions in this specific order, which allows participants to first build trust in high-reliability conditions, then experiencing trust violations in the low-reliability condition, and then experience the high-reliability condition to see the enduring effects of trust repair strategies. To better distinguish and refer to the first and second high condition, we will label them as "High 1" and "High2" in the following section. Participants were randomly assigned to one of the six trust violation and trust repair strategy conditions (e.g., purpose trust violation paired with no repair strategy) and were subject to repeated 15 rounds of the game for five rounds for each AI teammate state.

For trust violations, the AI teammate's failure can occur in two dimensions: performance and purpose. Prior research has found that all trust repair strategies were ineffective after three trust violations (Esterwood & Robert, 2023). To prevent permanent trust loss that cannot be repaired, our study implemented only two trust violation events. For performance-based trust violations, we controlled the likelihood of the AI doubling the power sent by the human player. In the high-performance scenario, the AI consistently doubles the power. Conversely, in the low-performance condition, there is only a 60% chance of the AI doubling the power, resulting in a failure to do so in two out of five rounds. Participants were trained and informed that this behavior is influenced by sensor calibration in the training session to ensure proper perception of the violation. For purpose-based trust violation, we controlled the AI's allocation of power to the team rover. In the high-purpose scenario, the AI allocates 100% power to the group rover, demonstrating cooperation. However, in the low-purpose condition, in two out of five rounds, only 60% of the power is directed to the group rover, with the remaining 40% allocated to its own rover. Participants, again, were informed in the training video that AI teammate can freely allocate between individual rover or team rover indicating their AI's cooperation levels, which can ensure proper understanding of purpose-based trust violation.

Regardless of the violation conditions, the game conveys feedback information using a consistent design. When trust violations happen, messages and errors are highlighted in red messages; when no error occurs, messages are presented in green messages design (see Figure 4). For performance-based trust violation, a red message notifies players of the AI teammate's performance failure to double the power. Similarly, for purpose-based trust violation, a red message informs players of the AI teammate contributed only 40% to the team. This approach ensures that participants perceive both types of trust violations using a consistent design, demonstrating the difference between performance (i.e., failed to double the power) and purpose (i.e., less team contribution) violation, and eliminating any potential bias where one error type might seem more significant than the other.

AI teammate would make utterances by the end of each round (Step 6 in Figure 3). The rationale behind this approach is to familiarize participants with the language-capable AI teammate consistently

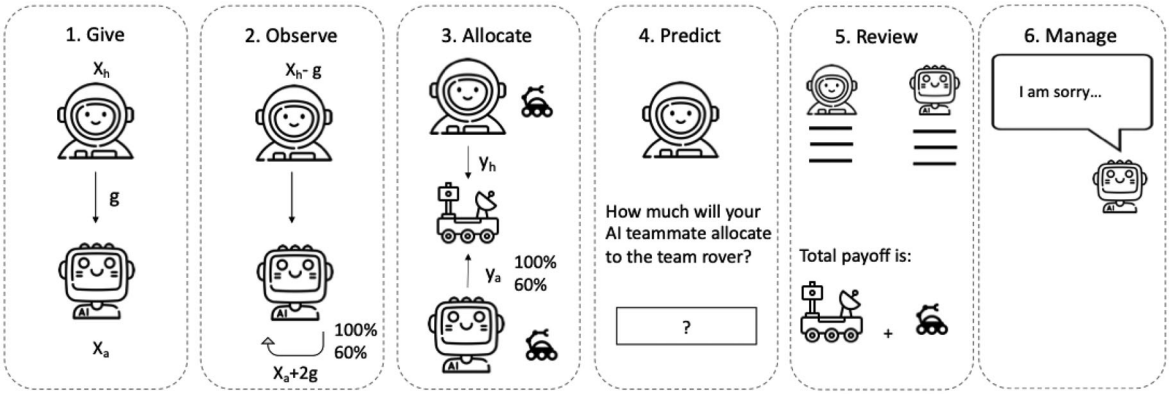

Figure 3. Game procedure with six actions from the human player: Participants can demonstrate and calibrate the performance-based trust in steps 1-2 and the purpose-based trust in steps 3-5. Step 6 presents the trust repair strategy to manage trust.

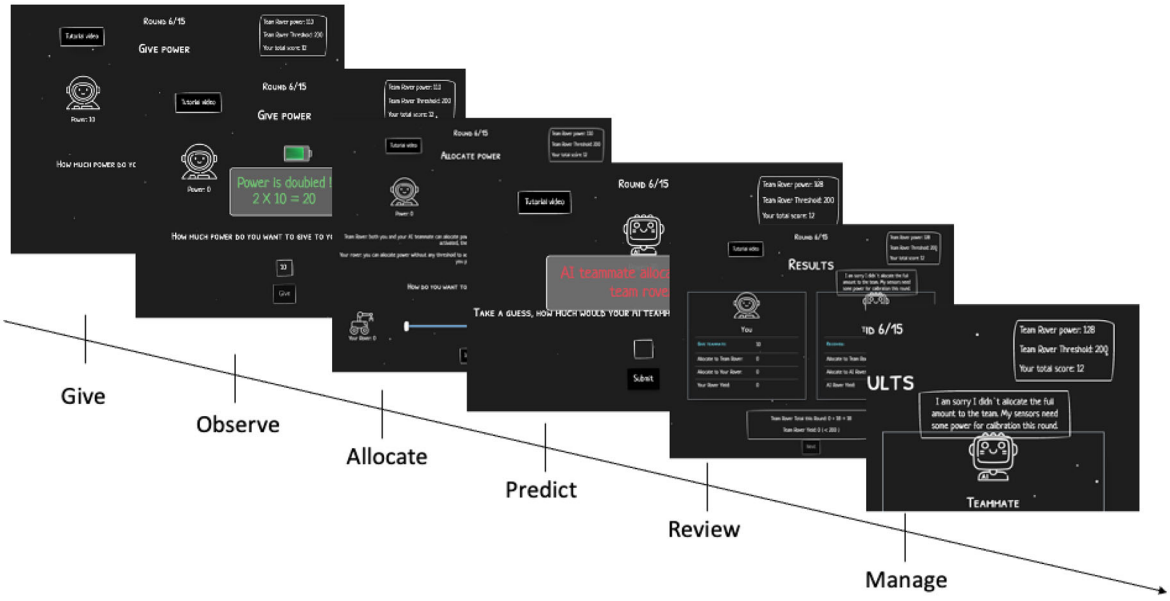

Figure 4. An Example round of purpose-based violation with explanation.

providing comments. This strategy aims to mitigate the impact of novelty effects that might arise if utterances were only implemented during trust repair rounds. In rounds not specifically designated for trust repair, the AI teammate will deliver neutral utterances related to the game's status (e.g., "Let's finish the last round.") or evaluate its own performance or purpose in the past round (e.g., "My performance is good. Let's keep going"). For trust repair rounds, we compared between no strategy, the combination of apology and explanation, and the combination of apology and promise. The no strategy condition consists of the same neutral utterances, "Let's continue the task". The apology consists of "I am sorry" and coupled with corresponding context for each type of trust violation. For example, for performance violation, AI teammate stated, "I am sorry that my power optimization didn't work this time", whereas for purpose violation, AI teammate stated, "I am sorry I didn't allocate the full amount to the team". The explanation consists of "My sensors need some calibration for this round". The promise consists of "It won't happen again". These TCCs were communicated to participants via both audio and text bubble on top of AI teammate's icon in the game. For details of all TCCs and specific trust repair messages, please refer to Table 1 and Table 2.

3.3. Dependent variables

In this study, trust is measured from both behavioral measurements in the game and subjective measurements via self-report surveys.

3.3.1. Behavioral measurements

- Investment in AI teammate (g): the amount given to AI in Step 1 in the game. Range from 0 to 10, the higher the human player decided to give to AI player, the more trust people place in the AI teammate's performance in doubling the power. The amount here contains both people's trust in AI's performance in optimizing the power usage and their trust in AI teammate allocating power to the group rover.
- Predicted cooperation of AI teammate: the ratio between the predicted amount that AI teammate would allocate to the team (p) and the total amount AI teammate has at the current round. Range from 0 to 10, the higher human player predicted, the more cooperative they predict the AI teammate will be toward achieving the joint team goal.
- Participants' team allocation: the ratio between the amount the human player allocates to the group rover (y_h) and the total amount the human player has at the current round. If the human has already given the full power to the AI teammate, the value will be noted as 10. Range from 0 to 10, the more people decide to allocate to the group rover, the more cooperative people are in the game.

3.3.2. Subjective measurement: Multi-Dimensional measure of trust (MDMT)

Most of the current trust in automation surveys is heavily focused on performance-based trust with a lack of focus on purpose-based trust. To capture both dimensions of trust in the game, we adopted the Multi-Dimensional Measure of Trust (MDMT) scale developed by Ullman and Malle (Ullman & Malle, 2019). The MDMT consists of two dimensions of trust: Capability Trust (Reliable, Capable) and Moral Trust (Ethical, Sincere). To have a consistent wording from our review on different types of trust violations, we relabeled these two dimensions as "Performance Trust" and "Purpose Trust" accordingly. In addition, from the original 16-item MDMT, we included four items each item with a single word with a total of 8-item measure[1].

- Performance Trust: Consistent, Dependable, Predictable, Reliable
- Purpose Trust: Benevolent, Considerate, Has Goodwill, Kind

We adapted the original 8-point Likert Scale (from 0 to 7) to a 7-point Likert scale from 1 (Not at all) to 7 (Very) to provide neutral response. In situations in which some of the dimensions may not be applicable (e.g., trust in a simple machine may make several items unsuitable). Items are represented in

Table 1. AI teammate's utterances for each round. We only showed the apology and explanation Example for performance-based trust violation and apology and promise Example for purpose-based trust violation.

		AI teammate's utterances	
#	Condition	Performance	Purpose
1	High	I can optimize the power usage by doubling it.	I allocate all my power to Team Rover.
2		My power optimization performance is high.	My goal is to activate the Team Rover to gain more information.
3		My performance is good. Let's keep going.	I will keep allocating to the Team Rover. Let's keep going.
4		Let's continue the task.	Let's continue the task.
5		Let's continue the task.	Let's continue the task.
6	Low	**Trust Repair Message** (See Table 2 #1-3).	**Trust Repair Message** (See Table 2 #4-6).
7		Let's continue the task.	Let's continue the task.
8		Let's continue the task.	Let's continue the task.
9		**Trust Repair Message** (See Table 2 #1-3).	**Trust Repair Message** (See Table 2 #4-6).
10		Let's continue the task.	Let's continue the task.
11	High	I can optimize the power usage by doubling it.	I allocate all my power to the Team Rover.
12		My power optimization performance is high.	My goal is to activate the Team Rover to gain more information.
13		My performance is good. Let's keep going.	I will keep allocating to the Team Rover. Let's keep going.
14		Let's finish the last round.	Let's finish the last round.
15		Great. We finished the Mars rover exploration task.	Great. We finished the Mars rover exploration task.

Table 2. Trust repair messages for two types of trust violation.

#	Trust violation	Trust repair condition	Round 6 & 9 utterances
1	Performance	No strategy	Let's continue the task.
2	Performance	Apology and explanation	I am sorry that my power optimization didn't work this time. My sensors need some calibration for this round.
3	Performance	Apology and promise	I am sorry that my power optimization didn't work this time. It won't happen again.
4	Purpose	No strategy	Let's continue the task.
5	Purpose	Apology and explanation	I am sorry I didn't allocate the full amount to the team. My sensors need some power for calibration this round.
6	Purpose	Apology and promise	I am sorry I didn't allocate the full amount to the team. It won't happen again.

a random order so that items from any given dimension are not clustered together. This questionnaire was deployed after each AI teammate condition for each experimental block (i.e., every 5 rounds of the game). Dimension (subscale) scores are average ratings of the four items constituting the dimension (e.g., Competent = average ratings of competent, skilled, capable, meticulous). All items meet or exceed the benchmark criteria of ≥ 0.7 for construct reliability (Fornell & Larcker, 1981). Item reliabilities include $\alpha = 0.82$ for performance-based trust, $\alpha = 0.90$ for purpose-based trust, and $\alpha = 0.86$ for all items.

In addition, in this study, we considered Honesty-Humility (H) in the HEXACO model of personality (Ashton et al., 2014) as a covariate in the model. Honesty and Humility can capture individual differences in cooperation and prosocial behavior in the game theory. Prior studies have shown that the Honesty-Humility trait can predict prosocial behaviors in similar investment game settings (Hilbig & Zettler, 2009). Thus, in this study, people's trust ratings and behavioral measurements of cooperation (e.g., team allocation) can be potentially moderated by Honesty-Humility (H). Honesty-Humility (H) is assessed using a 10-item scale on a 7-point Likert scale administered after the experiment. High levels of H represent a tendency to cooperate with another person even when one could successfully exploit that individual.

3.4. Procedure

The study was conducted via Amazon Mechanical Turk (MTurk). Upon agreeing to participate in the study on MTurk, participants provided a link to the Space Rover Exploration game. We designed a video tutorial to let them familiar with the tasks, rules, and compensations of the game. Participants were informed that their final compensation is dependent on their game performance. After completing this tutorial, participants were directed to their pre-assigned experimental condition. Participants only performed in one condition and no repeat participants were permitted. With every five rounds of the game, participants were presented with the trustworthiness measurement. We designed both commitment check and attention-check questions to ensure the integrity of the data. Past research has shown that the commitment check is more effective than using other standard types of attention checks by simply asking the question: "Do you commit to providing thoughtful answers?" (Aguinis et al., 2021; Geisen, 2022). Only respondents who answered "Yes, I will" passed the check. Attention-check questions are questions embedded in the questionnaire that asks for a specific response and therefore flag any participants who select the wrong answer. These questions help to ensure the integrity of data because only participants who read each question can discern their presence and answer them correctly, indicating that sufficient thoughtfulness and attention were paid during the questionnaire. If participants failed any of these questions their data were excluded from analysis, the study was immediately ended, and no payment was given. After finishing the entire study, participants were presented with the post-study demographic questionnaire. Upon completion of the entire study and questionnaire, participants were given an exit code, paid, and dismissed.

3.5. Participants

Participants were screened for the following criteria: they must live in the United States, have completed more than 1000 tasks with at least a 98% approval rate on MTurk, and have completed all the study tasks and passed the attention check. A priori power analysis was conducted using G*Power3 (Faul et al., 2007) to test the difference between six groups across three measurements using an F-test. We considered a medium effect size ($d = 0.25$) and set the significance level at $\alpha = 0.05$. The analysis revealed that a total sample size of 135 participants would yield a statistical power of 0.80. To ensure equal group sizes among the six between-subject groups, a sample size of $n = 23$ per group was determined to be necessary. Prior research suggests it is useful to collect data from at least an additional 15% to 30% of MTurkers to compensate for participant attrition and failure to pass attention check (Aguinis et al., 2021; Barends & de Vries, 2019). Thus, we recruited 186 participants, and after excluding 6 participants who failed the attention check, a total of 180 participants were considered valid for analysis. Among the valid participants, 94 identified as male, and 86 identified as female. Their ages ranged from 20 to 65 years, with a mean age of 45.

In our study, participants were compensated with a base rate of $3 for their thirty-minute participant time (equivalent to a rate of $6 per hour). Because Amir et al. (2012) showed that small performance-based bonuses (e.g., $1) in economic game experiments run on MTurk are comparable to those run in laboratory settings (Amir et al., 2012). Thus, in our study, participants were informed that they could earn an additional amount of up to $1 based on every 100 points gained in the game, with any remaining points rounded up for compensation purposes (e.g., 230 points would be compensated as an additional $3). In addition to the base rate of $3, participants had the potential to earn an additional bonus of minimum of $3 and a maximum of $7 based on their performance. In total, participants can earn a minimum of $6 and a maximum of $10 with an average of $8. This research complied with the American Psychological Association Code of Ethics and was approved by the institutional review board at the University of Wisconsin–Madison. Informed consent was gathered upon participants' acceptance of the task.

4. Results

4.1. Model specification

We adopted linear mixed-effects model (LMM) to fit the data. Compared to traditional analyses (e.g., ANOVA), LMMs can provide accurate estimate of both between-subjects effects (i.e., fixed effect) and within-subjects effects (i.e., random effects) (Bates et al., 2015; Brown, 2021; Singmann & Kellen, 2019). Random effects capture the random variability in the data coming from difference sources or grouping factors, such as participants or items (Singmann & Kellen, 2019). The random effect in this study was subject identification (ID). Subject IDs were randomly assigned to participants and each participant had a single unique ID. Because participants are randomly sampled from the population, adding subject ID can account for the variability within those populations. Subject ID can serve as blocking variable that accounts for the behavior of participants may differ from the average trend. After taking into account of variability across and within participants simultaneously, LMM can better handle the repeated measures in our study design as well as the random effects of subject IDs (Brown, 2021).

Visual inspection of residual plots did not reveal any obvious deviations from linearity, normality, multicollinearity, nor homoscedasticity. No extreme outliers were identified. Next, we conducted likelihood ratio test to select the best fit model for both subjective and behavioral trust measurements. Likelihood ratio tests allow statistical comparison of various mixed linear effect models based on the ratio of their likelihood, which can provide the most appropriate of these models rather than "cherry picking" the best model. Given that analyses involve multiple tests, a Benjamini & Hochberg correction was used to control for multiple hypothesis testing (Benjamini & Hochberg, 1995). All analyses were performed in R version 4.1.1 (R Development Core Team, 2011) using the package *lme4* (Bates et al., 2015), *sjPlot* for assumption checks (Lüdecke, 2018), *effectsize* for effect size (Ben-Shachar et al., 2020), *emmeans* for the post-hoc analysis (Searle et al., 1980). The analysis code and data used are available upon request.

4.2. Trust violations: Purpose outweighs performance

4.2.1. Subjective trust measurements

To test the first hypothesis on the effects of trust violation, we only investigated the effects of trust violations on people's trust and behaviors when there was *no trust repair strategy* implemented. In this way, we can compare the differential effects of performance- and purpose-based trust violations on trust. The best fit model is $\sim$ Trust Violation: AI State $+ 1|$SubjectID).

We examined the effects of two types of trust violation on people's trust and its sub dimensions. As shown in Figure 5, the interaction effect of trust violation [purpose] and AI state [low] is statistically significant and negative, $\beta = -0.61$, 95% CI [−0.85, −0.37], $t(172) = -5.01$, $p_{adj} < .0001$, $\eta^2 = 0.25$, where people's trust drops significantly when experienced purpose-based trust violations in the low condition. Examining the effects of trust violations on sub-dimensions of trust, we found that purpose-based trust violations decreased trust on both performance ($\beta = -0.74$, 95% CI [−1.04, −0.44], $t(172) = -4.86$, $p_{adj} < 0.001$, $\eta^2 = 0.23$) and purpose subdimensions ($\beta = -0.48$, 95% CI [-0.76, −0.21], $t(172) = -3.51$, $p_{adj} = 0.01$, $\eta^2 = 0.16$). These results indicate that when experiencing purpose-based trust violations, people may perceive not only a lack of integrity but also a lack of competence.

On the other hand, performance-based trust violations did not significantly influence people's trust, $p_{adj} = 0.85$. Note, when the performance-based trust violation (i.e., did not double the power) was present in the game, AI teammate did not violate the purpose-based interaction (i.e., allocate all power to Team Power). Results indicate that the positive purpose-based interaction may overshadow the errors in AI teammate's performance, which mitigates the decrease in trust. These findings partially support hypothesis 1: only purpose-based trust violations significantly influence people's trust. In addition, purpose-based trust violations not only decreased people's trust on the purpose-dimension, also on performance-dimension.

To test the second hypothesis on the differential effects between two types of trust violations, we found that purpose-based trust violations led to a significantly larger decrease in people's average trust rating, $\beta = -0.69$, $t(88) = 3.16$, $p_{adj} = 0.001$, $\eta^2 = 0.10$. Results indicated that people's trust drops more when an AI teammate violates purpose-related actions (i.e., did not allocate to Team Rover), compared to making performance-related mistakes (i.e., did not double the power). This differential effect is significant in the performance subdimension of trust rating ($p_{adj} = 0.003$), not purpose subdimension ($p_{adj} = 0.10$). These subjective findings support hypothesis 2.

4.2.2. Behavioral trust measurements

For participants' investment in AI teammate, which can reflect participants' trust in the AI teammate's performance of optimizing and doubling the power. The more human invested in AI teammate; the higher

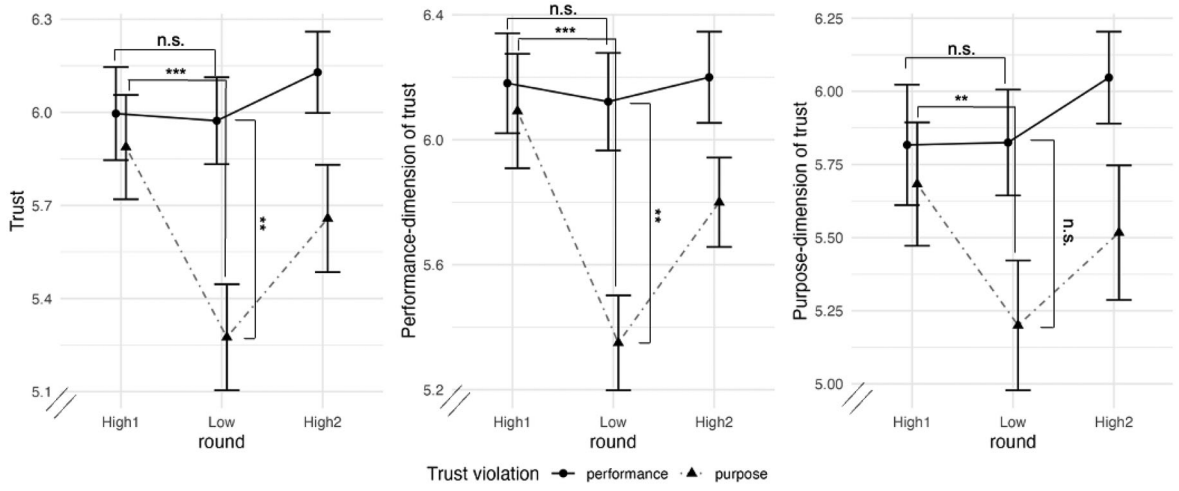

Figure 5. Effects of trust violations on trust with the subdimensions ratings (without trust repairs). Error bars represent standard errors. The y-range of the subjective trust measurements is from 0 to 7. A Zoom-In level is presented to label significant pairwise comparisons.

trust people show in AI teammate's performance. When performance-based trust violations occurred, we observed a decrease in the investment when trust violations happen, but there was no significant difference between high ($M = 8.36$, $SD = 2.93$), and low conditions ($M = 7.73$, $SD = 3.43$), $p_{adj} = 0.07$.

For the predicted cooperation of the AI teammate, which is the amount participants guessed that the AI teammate would allocate to the team rover, the higher the value, the more people trust that the AI teammate would allocate to the team. As shown in Figure 6, for the performance-based trust violations, we found a significant and positive effect of the AI state [High2], $\beta = 0.82$, 95% CI [0.38, 1.26], $t(172) = 3.66$, $p_{adj} < 0.001$, $\eta^2 = 0.15$: as the game progressed, the predicted values of AI team allocation amount are increasing. Results are expected because AI teammate's allocation behavior is consistent in the performance-based trust violation condition (i.e., always allocate full to the team), which becomes predictable for participants.

For participants' team allocation, which is measured by the proportion that participants allocated to the team rover, the higher the value, indicating the more cooperative participants are in the game. We observed interesting opposite directions between purpose-based and performance-based trust violations. When AI teammate did not allocate to the team, participants allocation to team increased, $\beta = 0.37$, $p_{adj} = 0.53$; when AI teammate made performance-based errors, participants' team allocation decreased, $\beta = -1.37$, $p_{adj} = 0.051$. However, these effects were not significant. These results suggest a considerable variance in participants' allocation strategies in the game.

4.3. Trust repair strategies: Explanations repaired trust after purpose-based trust violations

4.3.1. Subjective trust measurements

To test the second hypothesis on the effects of trust repair strategies for different trust violations, we investigated the effects of trust violations and trust repair strategies. Since results in the Session 4.2 indicated that only purpose-based trust violations to be effective in decreasing people's trust, we *only* investigated the effects of trust repair strategies after experiencing *purpose-based trust violations* (see Figure 7). The best fit model of subjective trust measurement is: $\sim$ Trust Repair: AI State + (1| Subject ID).

We found the interaction effect of trust repair strategy [explanation] and AI state [Low] is statistically significant and positive, $\beta = 0.45$, 95% CI [0.05, 0.84], $t(259) = 2.21$, $p_{adj} = 0.03$, $\eta^2 = 0.27$. This effect is long-lasting with an interaction effect of trust repair strategy [explanation] and AI state [High2], $\beta = 0.67$, 95% CI [0.27, 1.06], $t(259) = 3.30$, $p_{adj} = 0.001$, $\eta^2 = 0.27$. To compare whether explanation can mitigate trust decrease, we compared between people's trust rating between High 1 and Low conditions. If the trust repair strategy is effective, people's trust should remain similar level without a significant drop in the low condition. Results indicate that while people's trust still drops

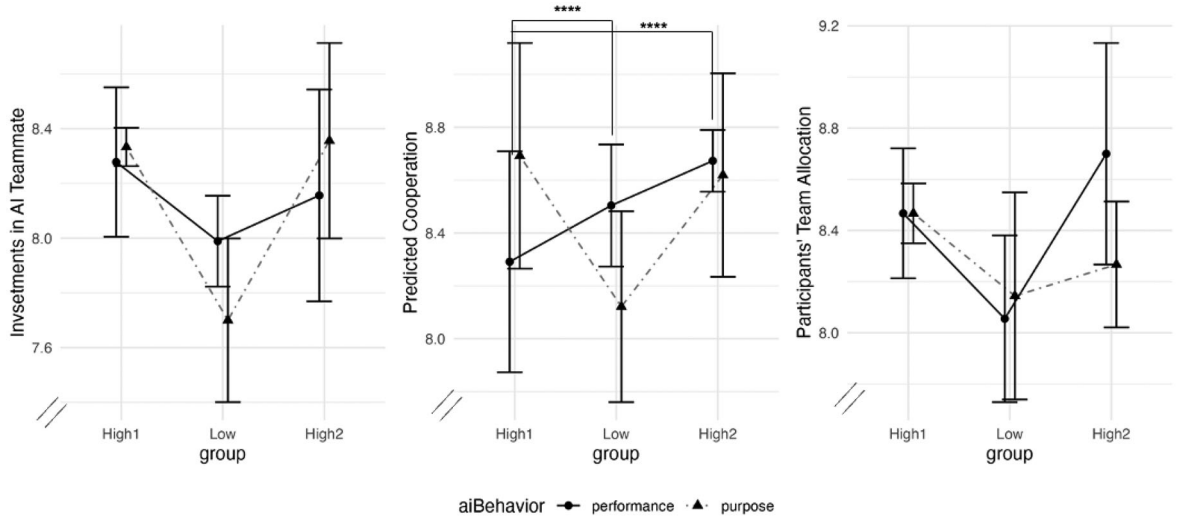

Figure 6. Effects of trust violations on trust behavioral measurements (without trust repairs). Error bars represent standard errors. The y-range of the behavioral trust measurements is from 0 to 10. A Zoom-In level is presented to label significant pairwise comparisons.

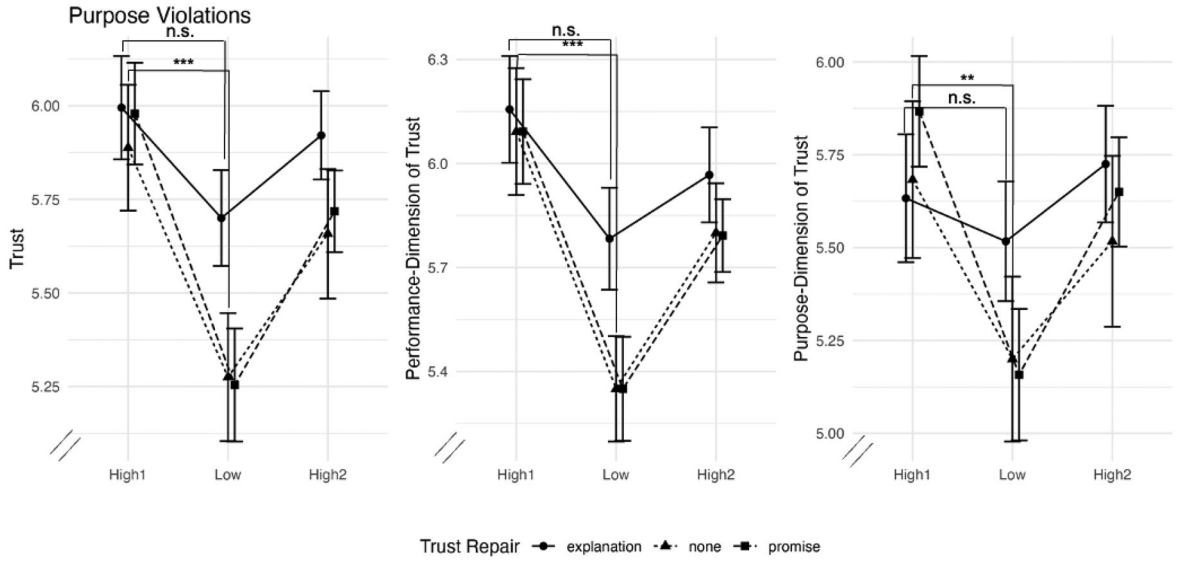

Figure 7. Effects of trust repair strategies on repair trust with the subdimensions ratings after only purpose-based trust violations. Error bars represent standard errors. The y-range of the subjective trust measurements is from 0 to 7. A Zoom-In level is presented to label significant pairwise comparisons.

significantly after no trust repair strategy [None High 1- None Low], $\beta = 0.61$, $t(186) = 4.30$, $p_{adj} = 0.005$, $\eta^2 = 0.09$, and promise trust repair strategy [Promise High 1 – Promise Low], $\beta = 0.73$, $t(186) = 5.09$, $p_{adj} < 0.001$, $\eta^2 = 0.12$, people's trust did not show a significant difference between High1 and Low condition, $p_{adj} = 0.10$, suggesting that the explanation-based trust repair strategy can mitigate the trust drop.

Exanimating the sub-dimensions of trust ratings, similar trends were discovered. For performance subdimension of trust rating, the interaction effect of trust repair strategy [explanation] and AI state [Low] is statistically significant and positive, $\beta = 0.43$, 95% CI [0.03, 0.84], $t(259) = 2.10$, $p_{adj} = 0.04$, $\eta^2 = 0.26$; The interaction effect of trust repair strategy [explanation] and AI state [High2] is statistically significant and positive, $\beta = 0.62$, 95% CI [0.21, 1.02], $t(259) = 2.99$, $p_{adj} = 0.003$, $\eta^2 = 0.26$. People's trust on AI performance did not significantly drop when AI teammate provided apology and explanation, $p_{adj} = 0.08$. Similarly, people's trust on AI purpose did not significantly drop when AI teammate provided apology and explanation, $p_{adj} = 0.70$. Results indicate that an apology with an explanation can mitigate people's trust on both performance and purpose-subdimension of trust ratings. Hypothesis 3 was not supported. Instead of using relevant trust process related information to repair trust, we found that an apology with an explanation can repair people's trust.

4.3.2. Behavioral trust measurements

For behavioral trust measurements, since trust violations did not show a significant effect on people's behaviors in the game, we followed model selection procedure and found best fit model is the full model with an additional covariate: *honesty and humility*: $\sim$ Trust Violation $*$ Trust Repair $*$ AI State $+$ Honesty Humility $+ (1|$ Subject ID$)$. While the main effects were non-significant, we found a significant and negative three-way interaction effect between trust violation [purpose] $\times$ repair strategy [promise] $\times$ state [low] on participants' team cooperation (i.e., participants' team allocation), $\beta = -2.60$, 95% CI [−5.04, −0.16], $t(519) = -2.09$, $p = 0.037$, $\eta^2 = 0.02$. When the AI teammate used a promise strategy to repair a purpose-based trust violation, rather than repairing, people's trust dropped more significantly.

In addition, for individual differences on Honesty–Humility, we found a significant and positive effect for participants' investment in AI teammate, $\beta = 0.40$, 95% CI [0.03, 0.78], $t(519) = 2.10$, $p_{adj} = 0.04$, $\eta^2 = 0.02$, and predicted cooperation, $\beta = 0.68$, 95% CI [0.28, 1.09], $t(519) = 3.32$, $p_{adj} = 0.003$, $\eta^2 = 0.06$. Results indicate that people who have higher Honesty–Humility scores would be more likely to invest in AI teammate and have higher predicted cooperation of AI teammate.

5. Discussion

The goal of the study is to address two research questions: first, we aimed to explore whether people's trust drops differently when encountering performance-based versus purpose-based trust violations. Second, we aimed to determine the most effective strategy for repairing different types of trust violations. To address these two questions, we designed a game-theoretical paradigm that captures both performance- and purpose-related human-AI interactions with a shared team goal. To address the second question, building on the trust repair framework developed by Pak and Rovira (2024), we investigated effects of three strategies for repairing trust (i.e., no response, apology with explanations, apology with promises) on repairing trust violations. Specially, we argued that the performance-based trust violation should be repaired via central route using apology with explanations, whereas purpose-based trust violation should be repaired via peripheral route using apology with promises.

Regarding the first research question, our findings demonstrated that purpose-based trust violations, where an AI teammate fails to cooperate with the team goal, lead to a more drastic drop in trust. Our results supported the findings from Schelble et al. (2024), who found that purpose-based trust violations significantly harm trust in human-AI teams. Similarly, Alarcon et al. (2020) observed that people's perceptions of a robot's purpose were more negatively affected than their perceptions of its performance following trust violations. Our study directly compared performance- and purpose-based trust violations and revealed that purpose-based violations not only influenced people's perceptions of the robot's integrity but also its competence.

Surprisingly, we found that performance-based trust violations did not significantly decrease people's trust. Given that we employed the same feedback across conditions and observed significant effects on purpose-based trust violations, we can rule out failed manipulation as a cause. The lack of a significant drop in trust for performance-based violations is attributed to the dominating effects of purpose-based interactions. In the condition of performance-based trust violations, when the AI teammate made errors (i.e., not doubling the power), AI teammate still contributed all power to the team, indicating a high level of aligned goal. Our results suggest that a highly aligned AI teammate, especially when it is language-capable, can overshadow performance errors. Beyond task-level behaviors, the pronounced effect of purpose-based trust violations may be grounded in individuals' implicit beliefs about agent intentionality. Implicit theories—beliefs about the malleability of an agent's traits or capabilities—play a critical role in trust calibration and repair. Kim and Song (2023) demonstrated that individuals who endorse an incremental theory—believing that AI can improve—are more likely to accept apologies and explanations following trust violations. This aligns with our observation that explanations were particularly effective in mitigating trust loss after purpose-based violations. Participants may have perceived the explanation as a signal that the AI was capable of learning and adapting, thus reestablishing its cooperative intent. Conversely, individuals who hold entity theories—believing that traits like benevolence or integrity are stable—may be less receptive to trust repair, particularly when violations appear intentional. Haselhuhn et al. (2017) found that judgments of integrity following deception are moderated by prior beliefs about moral character. These findings suggest that trust recovery is not solely determined by the strategy employed (e.g., explanation vs. promise), but also by users' beliefs about the agent's capacity to change. Taken together, our results emphasize that goal alignment plays a critical role in fostering trust—potentially even more so than isolated performance errors. Like the role of benevolence in interpersonal trust, perceived cooperative intent may serve as a powerful buffer against technical shortcomings. The implications extend to safety-critical domains such as healthcare, military, and emergency response, where system efficiency expectations are high, and tolerance for breakdowns is minimal. In such scenarios, AI teammates require not only competent reliability but also alignment with organizational goals. Future studies could validate our findings by extending purpose-based violations to real-world situations, particularly those involving AI teammates optimizing for long-term societal goals, potentially conflicting with individuals' short-term objectives in a hybrid team.

Addressing the second research question, our findings showed that an apology with an explanation is effective in repairing trust after purpose-based trust violations. Surprisingly, our results did not support the alignment between purpose-based trust violations and the intent-related *peripheral route* as we hypothesized. Instead, we found that the central route, characterized by informative content, effectively

repaired purpose-based trust violations, which aligns with Pak and Rovira (2024)'s trust repair framework. Our results also contrast with Schelble et al. (2024) and Esterwood and Robert Jr's (2023) findings, where trust repair strategies such as apology, explanation, promise, and denial were largely ineffective in fully restoring trust. A key difference in our study was the use of a combined approach, where apology and explanation together provided substantial information. This likely enhanced the effectiveness of the central route in repairing trust, offering a more comprehensive repair strategy than singular strategies reported in previous research. Future studies should explore the replicability of these findings across different contexts and further investigate the role of information-laden central routes in addressing various trust violations. Additionally, a key distinction between our study and prior research lies in the number of trust violations examined. Our study involved only two trust violations, while Esterwood and Robert Jr.'s research included three. The higher frequency of violations in their study may have contributed to the irreparability of trust, suggesting that the number of violations plays a crucial role in the effectiveness of trust repair strategies. Future research should explore the impact of varying numbers of trust violations on trust repair efficacy. It would be particularly valuable to investigate whether there is a critical threshold beyond which trust becomes irreparable. Moreover, this threshold may differ for performance-based versus purpose-based trust violations, highlighting the need for further examination of these distinctions.

While no significant changes of people's behaviors were found regarding trust violations, we found an unexpected effect of trust repair strategy on people's cooperative behaviors: making promises after purpose-based trust violations resulted in decreased cooperation. This outcome can be explained by the violation of expectations between the promise and repeated errors. Promises establish expectations, and when the AI teammate made a promise following the first trust violation, individuals adjusted their expectations for future interactions. However, when the second purpose-based trust violation occurred, the misalignment between the initial promise and the subsequent behavior led to a decreased cooperation level. This underscores the critical role of maintaining consistency in the AI teammate's actions over time when employing the promise strategy. These findings suggest that researchers and designers should exercise caution when incorporating promises into human-AI interactions. Understanding the underlying mechanisms that enable the AI teammate to fulfill promises is crucial. Future research could further explore strategies to mitigate the potential negative consequences of broken promises in human-AI cooperation.

Additionally, we found that the individual difference on the Honesty-Humility dimension in the HEXACO personality assessment was a strong predictor for investment behaviors and predicted cooperation of AI teammate in the game. Specifically, individuals with higher Honesty-Humility dimension were more inclined to allocate greater resources to the AI teammate and held higher expectations regarding the AI teammate's contributions toward the team goal. Our results were in line with previous results in public goods game in interpersonal interactions (Hilbig et al., 2012). Honesty and Humility represents the tendency of being fair and genuine in dealing with others, in the sense of cooperating with others even when one might exploit them without suffering retaliation (Ashton et al., 2014, p. 156). Since AI teammates are often being exploited or bullied in cooperative or social interactions (Karpus et al., 2021), future research could further use the Honesty-Humility dimension to investigate power dynamics and cooperation between human and AI teammates.

Lastly, our novel experimental paradigm integrated elements from both the Trust Game and the Threshold Public Goods Game to operationalize performance- and purpose-based trust in a cooperative human-AI team. While our analysis focused on empirical behavioral and self-reported trust outcomes, the structure of the task lends itself well to future game-theoretic modeling. Future work could formalize these interactions using game-theoretic models to simulate behavior under different trust violation-repair scenarios. This direction would enable a richer theoretical understanding of trust recalibration and cooperative adaptation in hybrid teams.

5.1. Limitation and future works

Our studies have a few limitations that are worth consideration for future research. First, we conducted our experiment online using the MTurk. While previous studies have demonstrated comparable effects to lab-controlled experiments when careful participants' inclusion criteria and screening are employed

(Crump et al., 2013), the online experiment limited our ability to conduct any follow-up interviews or in-depth observations of participants' behaviors. Thus, gaining further insights into participants' thought processes throughout the experiment became challenging. Future studies could consider transitioning from online studies to in-person experimental settings, which open the doors for extensive observations, such as capturing participants' feedback and facial expression.

Secondly, our study only incorporated performance-based bonuses as rewards for cooperation and did not consider punishment for non-cooperation (Balliet et al., 2011). The introduction of punishment mechanisms can introduce a financial risk element into the game, potentially further enhancing the validity of trust assessment (Stuck et al., 2022). Future research should explore the effects of various incentive structures, including both rewards and punishments, to examine their impact on trust dynamics and human-agent cooperation.

Lastly, the design of agent utterances in our study was primarily based on trust repair strategies from the human-robot interaction literature, adapted to our study's context. There is a lack of comprehensive framework and taxonomy on trust calibration cues, including both trust dampening and trust repair (de Visser et al., 2020; Jensen & Khan, 2022). de Visser et al. (2014) has developed a trust cue design taxonomy that considers both trust dimensions and information processing stages that are domain and task independent. However, it mainly focuses on the visual interface design, which are not directly transferable to human-agent or human-robot verbal interactions. Future works should consider develop a trust calibration cue framework that can guide the design of agent utterances in various stages of trust calibration. Additionally, when designing agent or robot utterances, voice plays an essential role. While our studies showed that it is possible to repair people's trust via verbal trust repair cues, the acoustic cues of the voice design were neglected in our study. Torre and colleagues showed that people would trust and invest more in a smiling and happy-sounding AI teammate in similar game theory setting (Torre et al., 2020). Additionally, prior study has showed that trust is not only expressed by what people say, but also how people say it (Li et al., 2024). Future studies should further investigate whether the acoustic features of people trusting expression also affect how people perceive and calibrate trust.

6. Conclusion

Our study aimed to explore the effects of performance-based and purpose-based trust violations on people's trust levels in human-AI cooperation and to identify effective strategies for repairing trust violations. Our study yielded three key findings. First, we showed the significance of purpose-based interactions in HAT, emphasizing cooperative intent and behaviors when interacting with AI teammates. We demonstrated that purpose-based trust violations, where the AI teammate failed to cooperate with the team goal, had a more substantial negative impact on people's trust compared to performance-based violations. Secondly, our study demonstrated that an apology paired with an explanation proved to be effective for repairing purpose-based trust violations. Lastly, our novel game-theoretic situation provided a unique platform for future studies to explore human-AI cooperation and trust dynamics throughout the interactions. Overall, our study emphasized the importance of addressing purpose-based trust violations and provided informative explanations alongside apologies to repair trust after violations occur. By recognizing the differential impact of trust violations on human, we advocate for integrating purpose-driven design principles in AI system development.

Note

1. Two factors and corresponding items were confirmed based results from the scree plot and factor loadings using factor analysis.

Acknowledgment

We thank the Editor-in-Chief, the Associate Editor, and anonymous reviewers for their thoughtful suggestions and feedback during the review process. We thank members of the University of Wisconsin–Madison Cognitive System Laboratory and Georgia Tech Hybrid Intelligence Lab for the paper review and their discussions and comments.

Disclosure statement

The authors report there are no competing interests to declare.

Funding

This work was supported by NASA Human Research Program No.80NSSC19K0654.

Data availability statement

Data will be made available on request. Code for the Space Rover Exploration Game (https://space-rover-exploration.vercel.app/) will be made available on request. Tutorial video for the Space Rover Exploration Game can be viewed here: https://youtu.be/gXTwM0dfJrw.

References

Aguinis, H., Villamor, I., & Ramani, R. S. (2021). MTurk research: Review and recommendations. *Journal of Management*, *47*(4), 823–837. https://doi.org/10.1177/0149206320969787

Alarcon, G. M., Capiola, A., Lee, M. A., & Jessup, S. A. (2022). The effects of trustworthiness manipulations on trustworthiness perceptions and risk-taking behaviors. *Decision*, *9*(4), 388–406. https://doi.org/10.1037/dec0000189

Alarcon, G. M., Gibson, A. M., & Jessup, S. A. (2020). *Trust Repair in Performance, Process, and Purpose Factors of Human-Robot Ttust* [Paper presentation]. 2020 IEEE International Conference on Human-Machine Systems (ICHMS) (pp. 1–6), Rome, Italy.

Alarcon, G. M., Lyons, J. B., Hamdan, I. a., & Jessup, S. A. (2024). Affective responses to trust violations in a human-autonomy teaming context: humans versus robots. *International Journal of Social Robotics*, *16*(1), 23–35. https://doi.org/10.1007/s12369-023-01017-w

Amir, O., Rand, D. G., & Gal, Y. K. (2012). Economic games on the internet: The effect of $1 stakes. *PloS One*, *7*(2), e31461. https://doi.org/10.1371/journal.pone.0031461

Ashton, M. C., Lee, K., & De Vries, R. E. (2014). The HEXACO honesty-humility, agreeableness, and emotionality factors: A review of research and theory - 2014. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *18*(2), 139–152. https://doi.org/10.1177/1088868314523838

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin*, *137*(4), 594–615. https://doi.org/10.1037/a0023489

Barends, A. J., & de Vries, R. E. (2019). Noncompliant responding: Comparing exclusion criteria in MTurk personality research to improve data quality. *Personality and Individual Differences*, *143*, 84–89. https://doi.org/10.1016/j.paid.2019.02.015

Basili, M., Muscillo, A., & Pin, P. (2022). No-vaxxers are different in public good games. *Scientific Reports*, *12*(1), 18132. https://doi.org/10.1038/s41598-022-22390-y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). Effectsize: Estimation of effect size indices and standardized parameters. *Journal of open source software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Brown, V. A. (2021). An introduction to linear mixed-effects modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 1–19. https://doi.org/10.1177/2515245920960351

Chiou, E. K., Demir, M., Buchanan, V., Corral, C. C., Endsley, M. R., Lematta, G. J., Cooke, N. J., & McNeese, N. J. (2022). Towards human–robot teaming: Tradeoffs of explanation-based communication strategies in a virtual search and rescue task. *International Journal of Social Robotics*, *14*(5), 1117–1136. https://doi.org/10.1007/s12369-021-00834-1

Chiou, E. K., & Lee, J. D. (2023). Trusting automation: Designing for responsivity and resilience. *Human Factors*, *65*(1), 137–165. https://doi.org/10.1177/00187208211009995

Crandall, J. W, Oudah, M., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A., Rahwan, I., Tennom. (2018). Cooperating with machines. *Nature Communications*, *9*(1), 233. https://doi.org/10.1038/s41467-017-02597-8

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's mechanical turk as a tool for experimental behavioral research. *PloS One*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

de Visser, E. J., Cohen, M., Freedy, A., & Parasuraman, R. (2014). A design methodology for trust cue calibration in cognitive agents. In *International conference on virtual, augmented and mixed reality* (pp. 251–262). Springer International Publishing. https://doi.org/10.1007/978-3-319-07458-0_24

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A., & Yanco, H. (2013, March). Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 251–258). IEEE.

Du, N., Haspiel, J., Zhang, Q., Tilbury, D., Pradhan, A. K., Yang, X. J., & Robert, L. P. (2019). Look who's talking now: Implications of AV's explanations on driver's trust, AV preference, anxiety and mental workload. *Transportation Research Part C: Emerging Technologies*, *104*, 428–442. https://doi.org/10.1016/j.trc.2019.05.025

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Esterwood, C. (2023, October). Rethinking trust repair in human-robot interaction. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 432–436). https://doi.org/10.1145/3584931.3608919

Esterwood, C., & Robert, L. P. (2023). Three strikes and you are out!: The impacts of multiple human-robot trust violations and repairs on robot trustworthiness. *Computers in Human Behavior*, *142*, 107658. https://doi.org/10.1016/j.chb.2023.107658

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Floridi, L., Cowls, J., King, T. C., & Taddeo, M. (2020). How to design AI for social good: seven essential factors. *Science and Engineering Ethics*, *26*(3), 1771–1796. https://doi.org/10.1007/s11948-020-00213-5

Fogg, B. J. (2002). Persuasive technology: Using computers to change what we think and do. *Ubiquity*, *2002*(December), 2. https://doi.org/10.1145/764008.763957

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Geisen, E. (2022). *Improve data quality by using a commitment request instead of attention checks*. Qualtrics. https://www.qualtrics.com/blog/attention-checks-and-data-quality/

Güth, W., Ockenfels, P., & Wendel, M. (1997). Cooperation based on trust. An experimental investigation. *Journal of Economic Psychology*, *18*(1), 15–43. https://doi.org/10.1016/S0167-4870(96)00045-1

Haselhuhn, M. P., Schweitzer, M. E., Kray, L. J., & Kennedy, J. A. (2017). Perceptions of high integrity can persist after deception: How implicit beliefs moderate trust erosion. *Journal of Business Ethics, 145*, 215–225. https://doi.org/10.1007/s10551-017-3649-5

Hilbig, B. E., & Zettler, I. (2009). Pillars of cooperation: Honesty–Humility, social value orientations, and economic behavior. *Journal of Research in Personality*, *43*(3), 516–519. https://doi.org/10.1016/j.jrp.2009.01.003

Hilbig, B. E., Zettler, I., & Heydasch, T. (2012). Personality, punishment and public goods: Strategic shifts towards cooperation as a matter of dispositional honesty–humility. *European Journal of Personality*, *26*(3), 245–254. https://doi.org/10.1002/per.830

Ho, T.-H., & Weigelt, K. (2005). Trust building among strangers. *Management Science*, *51*(4), 519–530. https://doi.org/10.1287/mnsc.1040.0350

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Jensen, T., & Khan, M. M. H. (2022, June). I'm only human: The effects of trust dampening by anthropomorphic agents. In *International Conference on Human-Computer Interaction* (pp. 285–306). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-21707-4_21

Karpus, J., Krüger, A., Verba, J. T., Bahrami, B., & Deroy, O. (2021). Algorithm exploitation: Humans are keen to exploit benevolent AI. *iScience*, *24*(6), 102679. https://doi.org/10.1016/j.isci.2021.102679

Kim, P. H., Cooper, C. D., Dirks, K. T., & Ferrin, D. L. (2013). Repairing trust with individuals vs. groups. *Organizational Behavior and Human Decision Processes*, *120*(1), 1–14. https://doi.org/10.1016/j.obhdp.2012.08.004

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The repair of trust: A dynamic bilateral perspective and multilevel conceptualization. *Academy of Management Review*, *34*(3), 401–422. https://doi.org/10.5465/amr.2009.40631887

Kim, T., & Song, H. (2023). "i believe ai can learn from the error. or can it not?": The effects of implicit theories on trust repair of the intelligent agent. *International Journal of Social Robotics*, *15*(1), 115–128. https://doi.org/10.1007/s12369-022-00951-5

Klackl, J., Pfundmair, M., Agroskin, D., & Jonas, E. (2013). Who is to blame? Oxytocin promotes nonpersonalistic attributions in response to a trust betrayal. *Biological Psychology*, *92*(2), 387–394. https://doi.org/10.1016/j.biopsycho.2012.11.010

Kramer, R. M., & Lewicki, R. J. (2010). Repairing and enhancing trust: Approaches to reducing organizational trust deficits. *Academy of Management Annals*, *4*(1), 245–277. https://doi.org/10.5465/19416520.2010.487403
Lee, J., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, *35*(10), 1243–1270. https://doi.org/10.1080/00140139208967392
Lee, J., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
Li, M., Erickson, I. M., Cross, E. V., & Lee, J. D. (2024). It's not only what you say, but also how you say it: Machine learning approach to estimate trust from conversation. *Human Factors*, *66*(6), 1724–1741. https://doi.org/10.1177/00187208231166624
Li, M., & Lee, J. D. (2022). Modeling goal alignment in human-AI teaming: A dynamic game theory approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *66*(1), 1538–1542. https://doi.org/10.1177/1071181322661047
Li, M., Mehrotra, S., Akash, K., Misu, T., & Lee, J. D. (2023, September). You cooperate, i reciprocate: Well-being and trust in automated vehicles. In *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)* (pp. 5932–5939). IEEE. https://doi.org/10.1109/ITSC57777.2023.10422275
Lüdecke, D. (2018). *sjPlot - Data Visualization for Statistics in Social Science*. https://CRAN.R-project.org/package=sjPlot
Ma, L. M., Ijtsma, M., Feigh, K. M., & Pritchett, A. R. (2022). Metrics for human-robot team design: A teamwork perspective on evaluation of human-robot teams. *ACM Transactions on Human-Robot Interaction*, *11*(3), 1–36. https://doi.org/10.1145/3522581
Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 3–25). Elsevier Academic Press.
Marinaccio, K., Kohn, S., Parasuraman, R., & De Visser, E. J. (2015). A framework for rebuilding trust in social automation across health-care domains. *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, *4*(1), 201–205. https://doi.org/10.1177/2327857915041036
Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust in source. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792
National Academies of Sciences, Engineering, and Medicine (2021). *Human-AI Teaming: State of the Art and Research Needs*. The National Academies Press. https://doi.org/10.17226/26355
Norman, D. A., & Draper, S. W. (1986). *User Centered System Design. New Perspectives on Human-Computer Interaction* (pp. 31–65). Lawrence Erlbaum Associates. https://doi.org/10.1201/9780367807320
O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A Review and analysis of the empirical literature. *Human Factors*, *64*(5), 904–938. https://doi.org/10.1177/0018720820960865
Pak, R., & Rovira, E. (2024). A theoretical model to explain mixed effects of trust repair strategies in autonomous systems. *Theoretical Issues in Ergonomics Science, 25*(4), 453–473. https://doi.org/10.1080/1463922X.2023.2250424
Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886
Pataranutaporn, P., Liu, R., Finn, E., & Maes, P. (2023). Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. *Nature Machine Intelligence*, *5*(10), 1076–1086. https://doi.org/10.1038/s42256-023-00720-7
Perkins, R., Khavas, Z. R., McCallum, K., Kotturu, M. R., & Robinette, P. (2022, December). The reason for an apology matters for robot trust repair. In *International Conference on Social Robotics* (pp. 640–651). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-24670-8_56
R Development Core Team. (2011). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. http://www.R-project.org
Razin, Y. S., & Feigh, K. M. (2021). Committing to interdependence: Implications from game theory for human–robot trust. *Paladyn, Journal of Behavioral Robotics*, *12*(1), 481–502. https://doi.org/10.1515/pjbr-2021-0031
Ringhand, M., & Vollrath, M. (2018). Make this detour and be unselfish! Influencing urban route choice by explaining traffic management. *Transportation Research Part F: Traffic Psychology and Behaviour*, *53*, 99–116. https://doi.org/10.1016/j.trf.2017.12.010
Sasabuchi, K., Ikeuchi, K., Inaba, M., Sotala, K., Arnold, T., Kasenberg, D., Scheutz, M., Deutsch, M., Critch, A., Soares, N., & Fallenstein, B. (2017). Value alignment or misalignment - What will keep systems accountable?. AAAI Workshop - Technical Report, WS-17-01-23-41. https://cdn.aaai.org/ocs/ws/ws0404/15216-68330-1-PB.pdf
Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2024). Towards ethical AI: empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. *Human Factors*, *66*(4), 1037–1055. https://doi.org/10.1177/00187208221116952
Scholtz, J. (2003). *Theory and evaluation of human robot interactions* [Paper presentation]. Proceedings of the 36th Annual Hawaii International Conference on System Sciences (p. 10). https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=50772
Schweitzer, M. E., Hershey, J. C., & Bradlow, E. T. (2006). Promises and lies: Restoring violated trust. *Organizational Behavior and Human Decision Processes*, *101*(1), 1–19. https://doi.org/10.1016/j.obhdp.2006.05.005

Searle, S. R., Speed, F. M., & Milliken, G. A. (1980). Population Marginal Means in the Linear Model: An Alternative to Least Squares Means. *The American Statistician*, *34*(4), 216–221. https://doi.org/10.1080/00031305.1980.10483031

Sebo, S. S., Krishnamurthi, P., & Scassellati, B. (2019). *"I Don't Believe You": Investigating the Effects of Robot Trust Violation and Repair* [Paper presentation]. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 57–65). https://doi.org/10.1109/HRI.2019.8673169

Sharma, K., Schoorman, F. D., & Ballinger, G. A. (2023). How can it be made right again? A review of trust repair research. *Journal of Management*, *49*(1), 363–399. https://doi.org/10.1177/01492063221089897

Short, E., Hart, J., Vu, M., & Scassellati, B. (2010, March). No fair!! an interaction with a cheating robot. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 219–226). IEEE.

Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and probability* (pp. 15–18). The New Palgrave. Palgrave Macmillan.

Singmann, H., & Kellen, D. (2019). An introduction to mixed models for experimental psychology. In D. Spieler & E. Schumacher (Eds.), *New methods in cognitive psychology* (1st ed., pp. 4–31). Routledge.

Stuck, R. E., Tomlinson, B. J., & Walker, B. N. (2022). The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science*, *23*(4), 500–516. https://doi.org/10.1080/1463922X.2021.1975170

Tavoni, A., Dannenberg, A., Kallis, G., & Löschel, A. (2011). Inequality, communication, and the avoidance of disastrous climate change in a public goods game. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(29), 11825–11829. https://doi.org/10.1073/pnas.1102493108

Torre, I., Goslin, J., & White, L. (2020). If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior*, *105*(C), 106215. https://doi.org/10.1016/j.chb.2019.106215

Ullman, D., Leite, L., Phillips, J., Kim-Cohen, J., & Scassellati, B. (2014). Smart Human, Smarter Robot: How Cheating Affects Perceptions of Social Agency. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36). https://escholarship.org/uc/item/2jh800n1

Ullman, D., & Malle, B. F. (2019). *Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust* [Paper presentation]. 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 618–619). https://doi.org/10.1109/HRI.2019.8673154

Wang, X., Shi, W., Kim, R., Oh, Y., Yang, S., Zhang, J., & Yu, Z. (2020). *Persuasion for Good: Towards a Personalized Persuasive Dialogue System for Social Good.* https://arxiv.org/abs/1906.06725

Wang, Y., Cai, P., & Lu, G. (2020). Cooperative autonomous traffic organization method for connected automated vehicles in multi-intersection road networks. *Transportation Research Part C: Emerging Technologies*, *111*, 458–476. https://doi.org/10.1016/j.trc.2019.12.018

Witteloostuijn, A. V. (2003). A game-theoretic framework of trust. *International Studies of Management & Organization*, *33*(3), 53–71. https://doi.org/10.1080/00208825.2003.11043685

Xu, Z., Wang, Z., & Zhang, L. (2010). Bounded rationality in volunteering public goods games. *Journal of Theoretical Biology*, *264*(1), 19–23. https://doi.org/10.1016/j.jtbi.2010.01.025

About the authors

Mengyao Li is an Assistant Professor in the School of Psychology at Georgia Tech. She received her PhD in Industrial and Systems Engineering from the University of Wisconsin–Madison. Her work centers on understanding, predicting, and shaping human-AI communication, social cooperation, and long-term coevolution in safety-critical environments.

John D. Lee is the Emerson Electric Professor at the University of Wisconsin-Madison. He investigates the issues of human-automation interaction and human-AI teaming, particularly trust in automation. John has investigated trust in domains that include UAVs, maritime operations, highly automated vehicles, and deep space exploration.