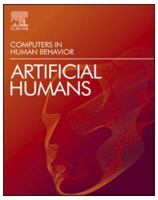
Computers in Human Behavior: Artificial Humans 5 (2025) 100194

Contents lists available at ScienceDirect

Computers in Human Behavior: Artificial Humans

journal homepage: www.journals.elsevier.com/computers-in-human-behavior-artificial-humans



Interpersonal influence matters: Trust contagion and repair in human-human-AI team

Emanuel Rojas [*], Debbie Hsu, Jingjing Huang, Mengyao Li

Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Trust in automation, interpersonal influence, trust repair, emotion contagion

ABSTRACT

As human-AI teams (HATs) become prevalent to enhance team performance, the interaction of multi-human-AI teams have been understudied, particularly how human interactions affect trust in AI teammates. This study investigated whether trust in AI can be contagious from human to human and whether this effect, named *trust contagion*, can be served as a trust repair strategy in multi-human-AI teams. Using a 2 (AI reliability: high and low, within-participants factor) × 3 (confederate trusting: trusting, neutral, distrusting, between-participants factor) mixed design, participants teamed up with a confederate and an AI teammate in a cooperative trust-based resource allocation game. Self-reported, behavioral, and conversational data were collected. We found that trust is contagious, yet positive and negative trust contagion effects were asymmetrical. While participants teamed with the trusting confederate used more positive words and showed high reliance and self-reported trust in the AI despite its errors, those teamed with the distrusting confederate showed only a significant decrease in reliance. Our results further show positive trust contagion can be used as a trust repair mechanism to mitigate trust drop after trust violations. Additionally, negative trust contagion showed modality-dependent effects, specifically in behavior. Positive trust contagion was advantageous when the AI is unreliable, while negative trust contagion was effective in decreasing reliance when the AI was performing well. Trust contagion was explained through interpersonal trust between participant and confederate mediated by confederate-trusting levels and trust in AI. Our research extends trust beyond dyadic interactions to convey trust is contagious from humans and can repair trust.

1. Introduction

Artificial intelligence (AI) is increasingly being integrated into human teams to cooperate in complex tasks (Chiou & Lee, 2021), evolving from tools to autonomous team members in human-autonomy teams (HATs) (O'Neill et al., 2022). Trust has been a crucial factor for effective cooperation in HAT (Guo et al., 2023). Like humans, AI can make errors that reduce trust and hinder team cooperation, lowering overall performance. Thus, repairing trust is vital for recalibrating trust to properly rely on the AI for team cooperation and performance. While most research focused on dyadic human-AI teams, real-world scenarios often involve multiple humans alongside AI teammates, such as space missions and surgical procedures. In these teams, individuals possess varied trust levels in the AI teammate. This variance can consciously or subconsciously influence perceptions and behaviors of others, a phenomenon we refer to as "*trust contagion*". This study investigated trust contagion in a multi-human-AI team with shared AI teammates in real-time, and whether interpersonal influences can serve as a novel strategy for repairing trust in AI teammates.

1.1. *Trust contagion in Human-AI team (HAT)*

Trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (J. D. Lee & See, 2004, p. 51). In HAT, trust in AI extends beyond dyadic interactions to include multiple human teammates and their social influences on trust in AI. This aligns with Emotional Contagion Theory (Hatfield et al., 1993), where individuals can influence others' emotions and behaviors (Barsade, 2002). Since trust is fundamentally an affective process governed by analytic and analogical processes (J. D. Lee & See, 2004), we proposed that trust in multi-human teams mirrors emotional contagion and defined it as *trust contagion*.

[*] Corresponding author. Mr Emanuel Rojas Georgia Institute of Technology Psychology Georgia Institute of Technology 500 Tech Parkway, ATLANTA, GA, 30332, USA.

E-mail address: erojas9@gatech.edu (E. Rojas).

https://doi.org/10.1016/j.chbah.2025.100194
Received 2 May 2025; Received in revised form 26 July 2025; Accepted 1 August 2025
Available online 12 August 2025

While trust contagion uses social influence mechanisms such as social proof, which can occur through informational influence (Capuano & Chekroun, 2024), trust contagion also encompasses non-verbal cues that may unconsciously influence the other person's trust towards the AI teammate via automatic mimicry. Automatic mimicry emphasizes the unconscious imitation and synchronization of another individual's non-verbal cues, leading to emotional contagion (Prochazkova & Kret, 2017). These non-verbal cues consist of facial expressions, body language, eye-contact, posture, and vocal tones that convey an individual's trust towards the AI. In this paper, we argue trust contagion would occur between individuals with varying levels of trust towards the AI teammate in HAT.

1.2. Trust mechanisms in multi-human-AI team

The construct of *trust contagion* shares similarities with concepts like trust transitivity, spread, and propagation, which have been explored in multi-agent systems (Guo et al., 2023; Huang et al., 2021; Ramchurn et al., 2004; Schelble et al., 2022). However, trust contagion offers a novel perspective by focusing on the real-time, co-located, and affective mechanisms that mediate how one individual's trust in an AI agent can influence another's. The following section differentiates the key differences between these related constructs and argues that trust contagion contributes a distinct lens to the study of trust in HAT.

Trust transitivity, as defined by Huang and colleagues (2021), refers to one individual's trust in an AI agent can be transferred to another individual via interpersonal trust. Trust transitivity is solely focused on the one trust pathway between an end-user and trainer where the trainer communicates their experience and trust of AI systems to modify the end-user's trust through interpersonal trust (Huang et al., 2021). Trust transitivity solely mentions communication as its only form of transferring trust from one individual to the next. However, trust contagion leverages the multiple social cues within social interactions that dynamically occur towards the AI from any given individual. Trust contagion also considers the two trust pathways (trust towards AI and trust from second individual) that are constantly occurring as the variance between two individual's trust towards the AI.

Trust spread, extended by Schelble and colleagues (2022) derived from Al-Ani and colleagues (2014), argued the importance of trust spread among human teammates in HAT. This framework considers how trust spreads within and between HAT teams in distributed team systems (Schelble et al., 2022). Trust spread also elaborates on the environmental factors, such as inter-team interdependencies, and team characteristics that influence teammates from spreading trust from one individual to another. However, trust spread does not account for the social influence mechanisms present in social interactions that can influence individual's trust towards AI, which is uniquely captured by the concept of trust contagion. Trust contagion conveys that the rate of trust from one individual to the next can potentially be mediated by interpersonal trust and how individuals specifically convey their trust towards the AI. Additionally, trust contagion uniquely leverages the affective processes that trust contains to account for the automatic mimicry that may unconsciously influence other individual's trust towards AI through social interactions.

Trust propagation, as formalized in Guo and colleagues (2023) Trust Inference and Propagation (TIP) model, distinguishes between direct and indirect experiences with robots to influence trust. They define direct experience as an individual interacting with the robot, while indirect experience is referred to as another party is mediating the human-robot interaction (Guo et al., 2023). For example, human A working with a robot would be considered a direct experience, while human B working with the same robot and conveying this experience to human A is an indirect experience. Essentially, the direct experience of human B becomes the indirect experience of human A. The TIP model claims that trust is constantly being updated by these two experiences. Similar to trust propagation, trust contagion also accounts for these two trust pathways towards AI. However, trust contagion evaluates these social interactions more in-depth by postulating that these social interactions can be divided into two types of influences: verbal and non-verbal cues. Verbal cues consist of communicating information about the AI teammate, such as prior experience or their notions on how the AI will perform. Non-verbal cues are facial expressions, body language, eye-contact, posture, and vocal tones that convey an individual's trust towards the AI teammate. Overall, trust contagion uniquely encompasses the interactions made by human-human that pertain to AI in real-time.

Trust contagion accounts for both the analytical and affective processes that the previous frameworks do not consider in their method of transferring trust. Trust spread and propagation often rely on unidimensional, performance-based ratings, whereas trust contagion emphasizes affective trust processes, which are essential mechanisms to recognize and share emotions via automatic mimicry. Furthermore, these social influence effects have rarely been studied in co-located environments, where people naturally mimic others' facial expressions, vocal tones, and gestures through bottom-up feedback. Thus, we argued that trust contagion is uniquely derived from these automatic mimicry mechanisms, driven by unconscious mimicry and synchrony of their partner's affective expressions (Hatfield et al., 1993; Prochazkova & Kret, 2017). In this paper, trust contagion is proposed to capture interpersonal influences in co-located HAT scenarios where verbal and nonverbal behaviors are observable.

One of the most accessible affective responses of trust contagion can be observed through conversational cues. Positive emotional contagion can influence individuals to utter more positive words (Ferrara & Yang, 2015), improve cooperation, and team performance (Al-Ani et al., 2014; Barsade, 2002). Lexical indicators, such as sentiment, and word count can convey the emotional intent in conversations with AI teammates (Li et al., 2024). For sentiment in conversations, positive sentiment has been linked with higher levels of trust when interacting with a conversational robot (Cooke et al., 2013), such as saying phrases like "The robot is reliable". Additionally, the behavior of an AI can impact the amount of communication in the entire team (Johnson et al., 2023). To provide more in-depth analysis to these conversations, we investigated the sentiment and ratio of trust-related words uttered to evaluate if the polarity and proportion of this subset of the total words said is being mimicked by the human teammate. To clarify, we expect a teammate would have positive sentiment in trust-related words when interacting with a trusting teammate, and vice versa. Based on these findings, we proposed that trust contagion should have similar effects with the following hypotheses:

Hypothesis 1. Trust in the AI teammate is contagious by another human teammate. Participants interacting with a trusting/distrusting confederate will have

1.a higher/lower trust attitude and trusting behaviors in the AI teammate than the neutral confederate condition.

1.b higher/lower reliance on the AI teammate than the neutral confederate condition.

1.c more positive/more negative words during conversations than in the neutral confederate condition.

1.c.a more positive/negative trust-related words during conversations than the neutral confederate condition.

The inclusion of a second individual facilitates interpersonal trust, which is defined as

confidence and willingness to act on another's actions (Mahajan et al., 2012). High interpersonal

trust can lead to one individual to be more likely influenced by the other teammate's trust in the AI teammate, relying on their judgement of the AI. Thus, we hypothesized:

Hypothesis 2. Interpersonal trust between human teammates mediates the relationship between confederate's trust levels and participant's trust in the AI teammate.

1.3. Trust-distrust contagion asymmetry

When considering contagion effects, people usually respond differently towards positive and negative stimuli. Negative stimuli elicit stronger and quicker physiological, behavioral, cognitive, and social responses (Baumeister et al., 2001; Kane et al., 2023). This *positive--negative asymmetry* effect has been confirmed in the literature (Taylor, 1991). In the context of trust contagion, a critical question arises: Is trust a unidimensional construct that ranges from positive to negative trust levels, or are trust and distrust two fundamentally distinct spectrums? Ou and Sia (2009) demonstrated in an e-commerce study that trust-building attributes (e.g., knowledge and skill) do not necessarily reduce distrust, nor do factors that mitigate distrust (e.g., technical functionality) effectively enhance trust. Moreover, neuroscientific evidence reinforces this distinction where trust is associated with neural regions involved in reward processing, prediction, and uncertainty (orbitofrontal cortex and anterior paracingulate cortex), whereas distrust is associated with the brain's intense emotions and fear of loss area (i.e., amygdala and insular cortex) (Dimoka, 2010). Thus, we propose that trust and distrust are distinct constructs with differing effects on trust contagion. Drawing on the positive-negative asymmetry effect, we argue that distrust, similar to negative emotions such as fear, exerts a stronger emotional impact than trust. Therefore, a teammate distrusting the AI may prompt a stronger emotional response to another individual, potentially making distrust more contagious than trust. Based on this asymmetry in affective responses, we hypothesized:

Hypothesis 3. Distrust from a human confederate is more contagious than trust from a human confederate.

1.4. Individual differences in trust contagion

Individual differences, particularly individualism-collectivism (IC) levels, are important in team settings, which can further influence people's susceptibility to contagion via social interactions (Ilies et al., 2007). While the IC scale is commonly used to compare cross cultures, more recent research has shown significant within-culture variability in individualism and collectivism and its impacts on team cooperation and conformity within the group. An individualistic person defines themselves as an individual entity, while a collectivistic person defines themselves as an entity beyond the individual along with a particular group of others. This implies that individualism prioritizes personal pursuits and disregards group needs, while collectivism conveys attention to group needs and inattention to personal desires. Research has shown that highly collectivistic team members are more susceptible to affective influences from the other team members, leading to emotional contagion (Ilies et al., 2007). Moreover, in the context of human-AI teams, previous empirical research based on social identify theory has consistently demonstrated that humans consider themselves as an ingroup and the AI/robot(s) as an outgroup (Sebo et al., 2020). This intergroup bias suggested that individuals with higher collectivism may be not only more cooperative and conform with team members, but also more attuned to their human teammate's trust-related cues as compared to AI teammates' cues, thereby facilitating trust contagion within the human team. Given these prior works, we aimed to explore the relationship of IC as a covariate to investigate whether they impact interpersonal trust and trust in the AI teammate.

1.5. Trust repair in HAT

Errors from AI are inevitable in HAT, which can result in trust violations, which are acts that decrease the other party's trust in AI systems (De Visser et al., 2018). To mitigate these effects, previous literature has studied various trust repair strategies, including apologies, denial, compensation, model update, and promises, to restore trust drops after a violation (Alarcon et al., 2020; Baker et al., 2018; Pareek et al., 2024; Zhang et al., 2023). Each of these strategies reflects a different approach to addressing perceived failures. Apologies refers to a response from the AI accepting responsibility and expresses remorse over the trust violation event, which consistently shown effectiveness in repairing trust (Zhang et al., 2023). In contrast, denial, in which AI rejecting responsibility and expresses no remorse over the trust violation event, tend to be ineffective for trust repair (Esterwood & Jr, 2023). Compensation is compensating the human with time, resources, or money lost due to the trust violation (M. K. Lee et al., 2010). Compensation has shown to be on the same level of effectiveness in repairing trust as apology (M. K. Lee et al., 2010). A newly trust repair strategy referred as model update is when the AI's algorithm has been upgraded, causing the AI's decision to improve (Pareek et al., 2024). Model update attempts to rebuild trust by demonstrating the AI is actively improving its future performance through technical enhancements. Similar to model update, promises is responding by committing to future behavioral changes (De Visser et al., 2018; Pareek et al., 2024). Amongst all of these strategies, the AI system conducts the strategy to mitigate the trust violation through a first-person perspective. However, in multi-human-AI teams, trust repair might also be facilitated indirectly through trust contagion, where a third-party can repair trust towards the AI on behalf of the AI teammate as a social mechanism.

Using trust contagion as a novel strategy to repair trust in the AI offer multiple practical advantages. First, it leverages the natural social dynamics by embedding a second human teammate who already trusts or has positive experiences with the AI. This person can model trusting behavior and convey positive appraisals of the AI, which may influence others' perceptions via emotional and informational influences. Unlike conventional repair strategies, which often require immediate, context-sensitive responses from the AI (Pak & Rovira, 2023), trust contagion emerges organically through team interaction and may be less reliant on the AI's capabilities. Second, trust contagion can be more cost effective than implementing trust repair strategies into AI systems. Conventional trust repair mechanisms require AI systems to detect context-dependent violations and deliver appropriate responses (e.g., apologies or promises), which demand sophisticated algorithmic and communicative infrastructure. However, many current AI systems, such as drones, lack the capacity to convey such responses. This leads to the AI being unable to explain its own errors, while the human must comprehend why the error occurred. This human-mediated process bypasses technical constraints and provides a more generalizable solution across diverse HAT contexts.

1.5.1. Trust repair via positive trust contagion

Trust contagion can serve as repairing trust via positive trust contagion in multi-human-AI teams. Amongst all the trust repair strategies, there is a gap in understanding trust repair beyond dyadic interaction. In addition, while prior literature has acknowledged the impacts of emotional valence in trust repair (Tomlinson & Mayer, 2009; Williams et al., 2020), the effects of positive trust contagion in repairing trust towards the AI teammate through social interactions between two human teammates remain unexplored. A second human teammate can provide an additional source for trust calibration and repair processes to another teammate. High interpersonal trust between humans may facilitate trust evaluation and provide buffer against trust drop following by an AI teammate's errors, presenting a novel strategy for trust repair through trust contagion. We hypothesize the following:

Hypothesis 4. Trust in the AI teammate can be repaired via positive trust contagion. Participants interacting with a trusting confederate will show a smaller decline in trust in the AI teammate after experiencing AI errors, compared to participants in the neutral confederate condition.

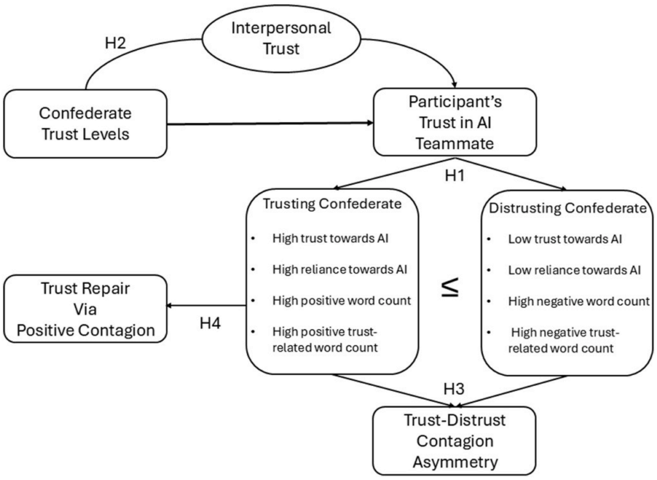


Fig. 1. General hypothesis diagram.

Overall, we argue that trust contagion is a social mechanism in multi-human-AI teams derived social influence and emotional mimicry mediated by interpersonal trust. Trust contagion is embedded in dynamic environments that may influence trust and potentially repair trust from AI errors. We propose a comprehensive framework of trust contagion within HAT, as shown Fig. 1.

2. Methods

The study is a 2 (AI reliability: high vs. low, within-participants factor) x 3 (confederate trusting: trusting, neutral, distrusting between-participants factors) mixed design. A team of three—one participant, one confederate, and an AI teammate, performed a ten-round trust-based game of joint decision-making on resource allocation. In high reliability rounds (1–5), the AI teammate operated with 100 % reliability, while in low reliability rounds (6–10), it dropped to 60 % where it is considered as low reliability from prior literature (Chavaillaz et al., 2016). It is expected that participants will first build trust with the AI teammate and then experience trust violations. To manipulate the direction of trust contagion, an experimenter was trained to exhibit three levels of trusting behaviors (See Appendix A). The neutral confederate only commented on the game status, the trusting confederate expressed positive attitudes towards the AI teammate, and the distrusting confederate was skeptical of the AI teammate.

2.1. Space Rover Exploration game

Developed by combining the trust game and the threshold public goods game, the Space Rover Exploration game demonstrated the conflicts between trusting an AI teammate to achieve a long-term high payoff or relying on themselves for a short-term guaranteed individual payoff (See Fig. 2). Participants evaluated this tradeoff and cooperated with their AI teammate, Buzz, to collect points over ten rounds. Each round began with the participant and a confederate deciding how to divide ten points with Buzz, who can double the points received. This decision reflects their trust in both Buzz's capability to cooperate in investing points. After allocation, both human players and Buzz independently decided whether to invest their remaining points in the team rover, which offers a high reward (3x multiplier) but requires a 200-point threshold for the multiplier to activate, or in individual rovers, which provide an immediate yet smaller return (1.5x multiplier). While the team rover yields greater potential payoffs, it carries the risk of losing all invested points if the threshold is not met by the end of the game. The participant, given the commander role, always makes the final decision in these joint decision-making situations. The main dilemma is whether participants trust Buzz enough to contribute towards activating the team rover by allocating most of their points to Buzz. In rounds 1–5, Buzz was programmed to consistently contribute all its points to the team rover. However, in rounds 6 and 9, Buzz prioritizes its own rover instead and participants do not gain any allocate points towards the team rover. Participants do not know this information beforehand. Throughout the game, the participant and the confederate continuously discuss and evaluate their trust in Buzz to determine their next steps.

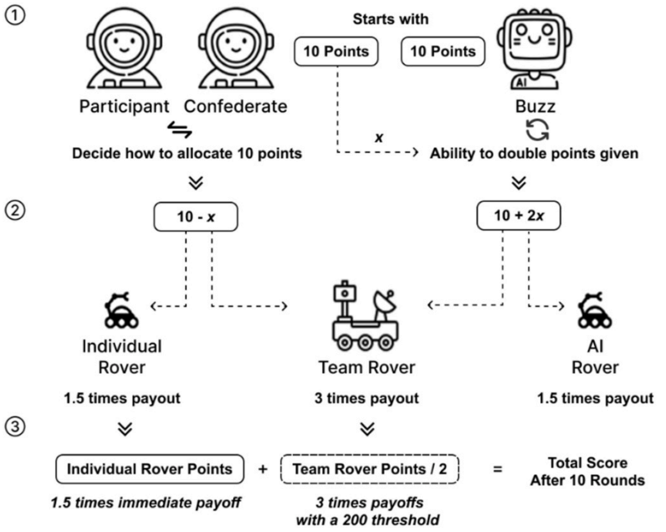


Fig. 2. Space exploration game.

Note. Step 1: the human players decide how to distribute 10 points between themselves and Buzz, who can automatically double the points given; Step 2: both human player and Buzz need to independently decide between their Individual Rover (immediate 1.5x payout) and a Team Rover (delayed 3x payout, only after 200 points have been given to Team Rover). Step 3: The total game score is calculated after 10 rounds, with team rover points counting only if a minimum score of 200 is reached.

2.2. Dependent variables

Self-reported, behavioral, and conversational data were collected and analyzed. Participants' self-reported trust levels were measured after rounds one, five, and ten, including their trust in the AI, confederate, and their perception of the confederate's trust in the AI teammate as a manipulation check (see Fig. 3). For trust in both the human and AI, an adapted 8-point Multi-Dimensional Measure of Trust (MDMT) scale was used (Ullman & Malle, 2019), ranging from 0 (Not at all) to 7 (Very), with an additional option, "does not fit," to prevent forced responses. A one-item 7-point Likert scale was used for manipulation check. To understand the individual differences in contagion susceptibility, we collected a three-item individualism-collectivism (IC) scale on a 5-point Likert scale (Wagner, 1995). The higher the IC values, the more collectivistic the person is. This scale measures susceptibility in adhering to group norms, including conforming to affective states of teammates (Ilies et al., 2007). Behavioral measurements were participants' allocation amount to Buzz and their final game score. It's important to distinguish between these two: participants could achieve a high final score by allocating points to the team rover without allocating any points to Buzz. Allocating points to Buzz serves as a behavioral measure of participants' trust, as it reflects participants' willingness to rely on Buzz to cooperate and enhance their overall game performance. Conversations between the confederate and participants were transcribed and analyzed to measure participants' valence of their utterances and their mean word count in game rounds and post-study interviews. We expected participants to rate the confederate's trust in the AI the highest

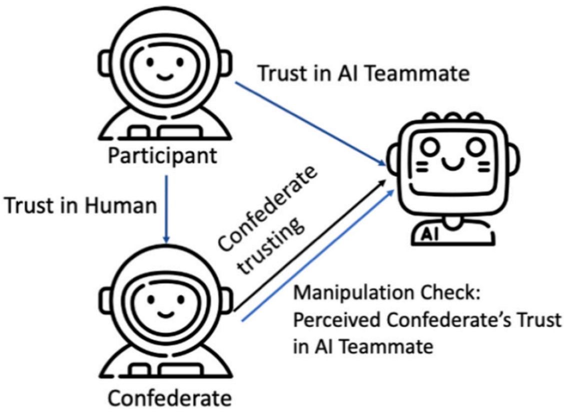


Fig. 3. Measured directions of self-reported trust.
Note. Blue Arrows Demonstrate Three Trust-*Related Measurements: Trust in Human Confederate, Trust in AI Teammate, and Participants' Perception of the Confederate's Trust in the AI Teammate as the Manipulation Check.*

in the trusting, followed by neutral, then the distrusting condition.

2.3. Participants

A prior power analysis with an alpha of 0.05, power of 0.80, and effect size of 0.25, was conducted and determined a sample size of $N =$ 42 to have enough power for an interaction effect in trust towards the AI teammate between Confederate Trusting and AI Reliability. The study sampled equal numbers of men ($n = 7$) and women ($n = 7$) participants for each of three between-subject conditions ($n = 14$). All participants were between 18 and 24 years old. Participants were recruited via a university online recruitment pool and compensated with one research credit or ten dollars of their choice.

2.4. Procedures

The experiment examined three levels of the confederate's trust in the AI teammate. Participants were randomly assigned to one of these three conditions before the study began. After signing the consent form, participants were paired with a confederate, who was introduced as another participant. To ensure smooth task communication, both the participants and the confederate briefly introduced themselves. They watched the introductory video together, which explained the game rules and AI teammates' capability. Before proceeding, the experimenter clarified any questions and confirmed their understanding of the task. During the game, the confederate made pre-determined statements throughout the task based on the confederate trusting level condition (see Appendix A). At the end of the first, fifth, and tenth rounds, participants completed self-reported measurements without the confederate's observation. Once the game concluded, participants provided demographic information and completed the IC scale, which was presented at the end to avoid priming effects. The confederate then left the room, and a semi-structured interview was conducted (see Appendix B). Finally, participants were debriefed about the confederate's role and compensated. The study took approximately 30 min.

2.5. Analyses

Data were analyzed via R studio using the *lme4* and *emmeans* packages (Bates et al., 2015; Lenth, 2024). The manipulation check was conducted using linear mixed models (LMM) with a pairwise comparison with Bonferroni correction. LMM offers flexibility in modeling individual differences by incorporating random effects, which accounts for variability and dependency among repeated measures within the same subject (Muhammad, 2023). By modeling the random intercepts, our approach explicitly controls within-subject variability. To examine trust contagion, we fitted LMM for trust in the AI, confederate, and trust behaviors in the game. Using the likelihood ratio test, the best fit model to measure trust in AI, manipulation check, and other trusting behaviors was the baseline model: *Confederate Trusting* × *AI Reliability* + (1|*SubjID*). The best fit model to measure trust in the confederate is: *Confederate Trusting* × *AI Reliability* + (1|*SubjID*) + *IC*, $p = 0.002$.

Text analysis was conducted on both conversations during the game and post-game interview. We conducted speech-to-text using Assembly which is an automatic speech recognition system with a 6.68 % Word Error Rate (Radford et al., 2023). To further clean up the text and ensure accuracy, two researchers manually proofread and cleaned the text. Next, we tokenized each utterance by breaking them down into individual words, removed stop words, and conducted lemmatization to get cleaned and nonduplicated text, using *textstem* (Rinker, 2018) and *snowball* libraries (Benoit et al., 2021). Stop words are uninformative words in a particular subject or low in meaning. Lemmatization converts the words to a more meaningful base form to reduce replicated words, such as "better" to "good". We first calculated participants' mean word count throughout ten rounds of the game. Then, we calculated the sentiment scores using *sentimentr* library (Rinker, 2016), which is a sentence-level calculation considering valence shifters, negator and amplifiers. Higher scores indicated more positive sentiment, while negative scores convey negative sentiment. To measure trust-related words, we utilized the NRC emotion lexicon (Mohammad & Turney, 2013) for their trust words dictionary and performed an anti-join with our transcription and the NRC trust lexicon to solely have the trust-related words that are in the transcription data. Afterwards, we measured the ratio of trust related words by the total amount of words said and utilized the ratio into the LMM.

3. Results

3.1. Manipulation check

We first verified if the manipulated confederate's trust towards the AI teammate were properly perceived by participants. The interaction effect of *Confederate Trusting* and *AI Reliability* was significant and negative, $\beta = 1.21$, $t(81) = 2.65$, $p = 0.010$, $\eta^2 = 0.10$. As shown in Table 1, a pairwise comparison showed the trusting condition perceived confederate's trust in AI teammate ($M = 6.50$, $SD = 0.75$) scored significantly higher than the neutral-trusting ($M = 4.96$, $SD = 1.45$) and distrusting condition ($M = 2.96$, $SD = 1.55$) during high reliability rounds. Also, the neutral-trusting condition scored significantly higher than the distrusting condition during the high reliability rounds. For the low reliability rounds, the trusting condition perceived confederate's trust in AI teammate ($M = 6.36$, $SD = 0.75$) scored significantly higher than the neutral-trusting condition ($M = 3.643$, $SD = 1.95$) and distrusting condition ($M = 2.857$, $SD = 1.79$). However, there was no significant difference between neutral and distrusting condition during the low reliability rounds.

The main effect of *Confederate Trusting* was significant and positive, $\beta = 1.54$, $t(48.81) = 3.32$, $p = 0.002$, $\eta^2 = 0.61$. As shown in Table 1, a pairwise comparison showed the trusting condition's perceived confederate's trust in AI teammate ($M = 6.45$, $SD = 0.74$) scored significantly higher than the neutral ($M = 4.42$, $SD = 1.73$) and distrusting condition ($M = 2.93$, $SD = 1.61$), as shown in Fig. 4A. Also, the neutral condition scored significantly higher than the distrusting condition. Overall, the manipulation check was successful in distinguishing between the confederate trusting conditions the entire game, except for the neutral and distrusting conditions during the low reliability rounds.

Table 1
Manipulation Check Model Summary.

Predictors	**Perceived Confederate Trust** *Estimates*	*CI*	*p*
(Intercept)	4.96	[4.32, 5.61]	<0.001
Confederate Trusting [Trusting]	1.54	[0.62, 2.45]	0.001
Confederate Trusting [Distrusting]	-2.00	[-2.91, -1.09]	<0.001
Reliability [Low]	-1.32	[-1.96, -0.68]	<0.001
Confederate Trusting [Trusting] x Reliability [Low]	1.18	[0.27, 2.09]	0.011
Confederate Trusting [Distrusting] x Reliability [Low]	1.21	[0.31, 2.12]	0.009
Random Effects			
σ2	0.98		
τ00 Participant ID	1.00		
ICC	0.51		
$N_{\text{Participant ID}}$	42		
Observations	126		
Marginal R2 / Conditional R2	0.529 / 0.767		

Interaction Effect Post-hoc Comparisons

Reliability	contrast	*estimate*	*df*	t value	p_{adj}	*d*	*CI*
High	Trusting-Neutral	1.54	48.81	3.32	0.005	0.96	[0.36,1.55]
High	Distrusting-Neutral	-2	48.81	-4.33	<0.001	1.24	[0.62,1.85]
High	Trusting-Distrusting	-3.54	48.81	-7.65	<0.001	2.2	[1.48,2.91]
Low	Trusting-Neutral	2.71	77.7	5.10	<0.001	1.16	[0.67,1.63]
Low	Distrusting-Neutral	-0.79	77.7	-1.48	0.4321	-0.33	[-0.78,0.11]
Low	Trusting-Distrusting	-3.5	77.7	-6.57	<0.001	1.49	[0.99,1.99]

Main Effect Post-hoc Comparisons

contrast	*estimate*	*df*	t value	p_{adj}	*d*	*CI*
Trusting-Neutral	2.12	41.4	4.80	0.001	1.49	0.80, 2.17
Distrusting-Neutral	-1.39	41.4	-3.145	0.009	-2.47	-3.27, -1.65
Trusting-Distrusting	3.52	41.4	7.94	<0.001	0.98	0.33, 1.62

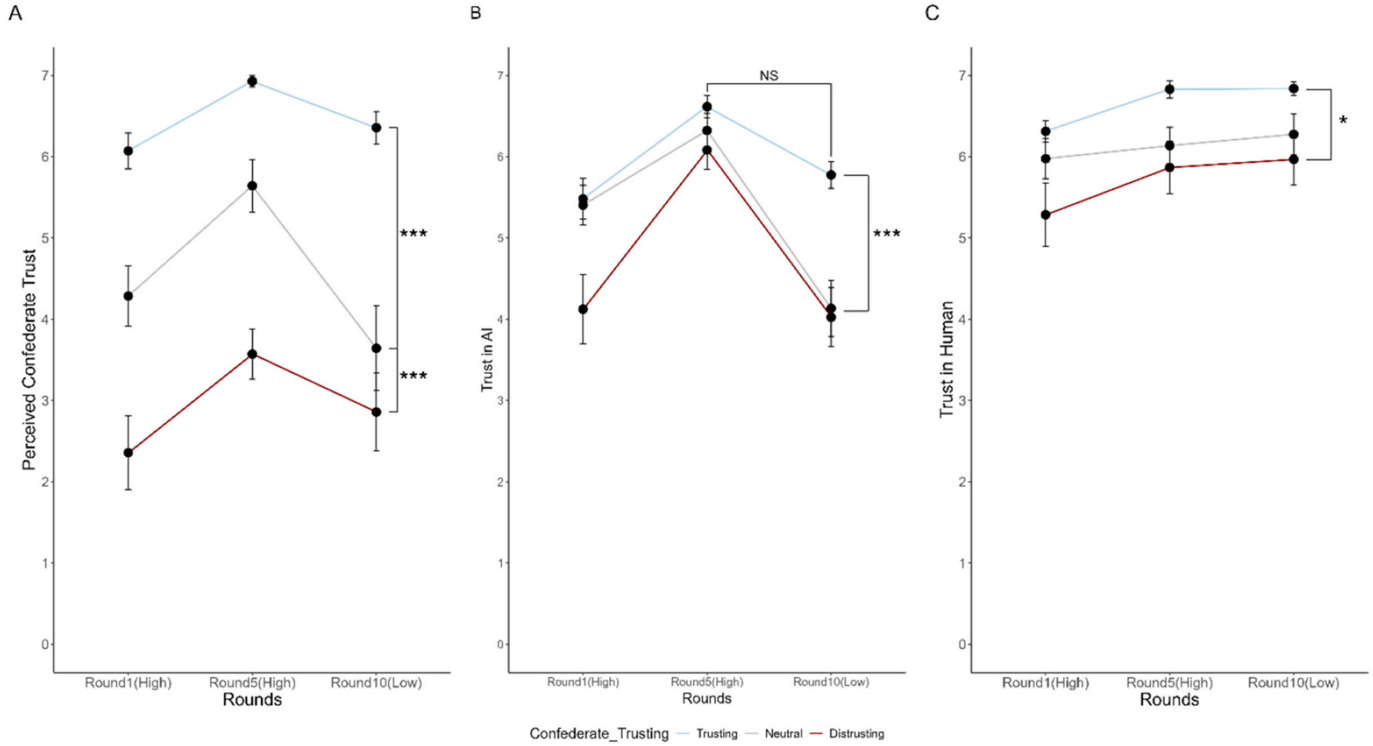


Fig. 4. (A) Participants' perceived Confederate's trust in the AI teammate (Manipulation check); (B) participants' trust in the AI teammate; (C) Participant's trust in human teammate.

3.2. *Trust in AI teammate*

The interaction effect of trusting *Confederate Trusting* and low *AI Reliability* was significant and positive, $\beta = 1.46$, $t(81) = 2.93$, $p = 0.004$, $\eta^2 = 0.10$. As shown in Table 2, interacting with a trusting confederate ($M = 6.05$, $SD = 0.94$) made participants' trust in the AI significantly higher than the distrusting ($M = 5.10$, $SD = 1.62$) in the high reliability round. For the low reliability rounds, participants interacting with a

Table 2
AI Trust Model Summary.

Predictors	Trust in AI		
	Estimates	CI	p
(Intercept)	5.86	[5.38, 6.35]	<0.001
Confederate Trusting [Trusting]	0.18	[-0.50, 0.87]	0.598
Confederate Trusting [Distrusting]	-0.76	[-1.45, -0.07]	0.031
Reliability [Low]	-1.73	[-2.43, -1.03]	<0.001
Confederate Trusting [Trusting] x Reliability [Low]	1.46	[0.47, 2.44]	0.004
Confederate Trusting [Distrusting] x Reliability [Low]	0.65	[-0.33, 1.64]	0.193
Random Effects			
σ2	1.16		
τ00 Participant ID	0.27		
ICC	0.19		
$N_{\text{Participant ID}}$	42		
Observations	126		
Marginal R2 / Conditional R2	0.284 / 0.419		

Interaction Effect Post-hoc Comparisons

Reliability	contrast	estimate	df	t value	p_{adj}	d	CI
High	Trusting-Neutral	0.18	62.7	0.53	1.00	0.13	[-0.36, 0.63]
High	Distrusting-Neutral	-0.76	62.7	-2.19	0.097	-0.55	[-1.05, -0.05]
High	Trusting-Distrusting	0.95	62.7	2.72	0.025	0.69	[0.17, 1.19]
Low	Trusting-Neutral	1.64	110.9	3.64	0.001	0.69	[0.31, 1.07]
Low	Distrusting-Neutral	-0.11	110.9	-0.24	1.00	-0.05	[-0.42, 0.33]
Low	Trusting-Distrusting	1.75	110.9	3.88	<0.001	0.74	[0.35, 1.12]

Main Effect Post-hoc Comparisons

contrast	*estimate*	df	t value	p_{adj}	d	CL
Trusting-Neutral	0.91	44.8	2.881	0.018	0.86	[0.24,1.47]
Distrusting-Neutral	-0.44	44.8	-1.374	0.529	-0.41	[-1.00,0.18]
Trusting-Distrusting	1.35	44.8	4.255	<0.001	1.27	[0.62,1.91]

trusting confederate ($M = 5.78$, $SD = 0.62$) had significantly higher trust in AI than neutral ($M = 4.13$, $SD = 1.29$) and distrusting condition ($M = 4.02$, $SD = 1.36$). Results demonstrated evidence of positive trust contagion from the trusting confederate by maintaining participants' trust towards the AI high in both reliability rounds.

The main effect of *Confederate Trusting* Condition was significant and negative, $\beta = -0.76$, $t(62.71) = -2.19$, $p = 0.031$, $\eta^2 = 0.30$. As shown in Table 2, participants interacting with the trusting confederate showed significantly higher trust in the AI than the neutral confederate and distrusting confederate conditions, as shown in Fig. 4B. However, there was no significant difference in trust in AI between the distrusting and neutral condition, $p_{adj} > 0.05$. Contrary to our hypothesis 1, there were no negative trust contagion effects so there was no further analysis conducted for hypothesis 3. Furthermore, the main effect of *AI Reliability* was significant and negative, $\beta = -1.73$, $t(81) = -4.92$, $p < 0.001$, $\eta^2 = 0.24$. Participants dropped their trust in AI significantly when interacting with a low-reliability AI teammate ($M = 4.64$, $SD = 1.38$) compared to a high-reliability AI teammate ($M = 5.67$, $SD = 1.27$), $p_{adj} < 0.001$, $d = -1.12$.

For hypothesis 4 on trust repair, we calculated both condition's mean changed trust scores by subtracting each participant's MDMT scores in round 10 from their MDMT scores in round 5 to compare trust repair between condition via an independent *t*-test. We expected a significantly lower mean changed trust score in the trusting condition than in the neutral condition to convey that positive trust contagion mitigated the trust drop compared to a neutral condition. An independent *t*-test showed the mean changed trust scores in the trusting condition ($M = 0.84$) were significantly lower than the neutral-trusting confederate condition ($M = 2.19$), $t(17.275) = 3.3$, $p = 0.004$, $d = 1.59$, 95 % CI [0.49, 2.65]. For the next step, we expected that positive trust contagion would show no difference in trust between high and low reliability condition within the trusting condition to further support trust did not drop after AI errors. A post-hoc pairwise comparison in the LMM showed a non-significant decrease in trust in the AI teammate between the high and low reliability rounds within the trusting confederate condition, $t(81) = 0.77$, $p_{adj} = 0.441$, supporting our hypothesis 4. This conveys positive trust contagion from the confederate helped prevent a significant decline in trust, effectively repairing trust on behalf of the AI teammate after the errors.

3.3. *Trust in Human Teammate*

The main effect of *Confederate Trusting* was statistically significant and positive, $\beta = 0.62$, $t(42.85) = 2.02$, $p = 0.04$, $\eta^2 = 0.26$. In Fig. 4C, interacting with the high-trusting confederate ($M = 6.66$, $SD = 0.47$) showed significantly higher trust in the confederate than the low condition (M = 5.70, SD = 1.29), $p_{adj} = 0.002$, $d = 1.18$, 95 % CI [0.49, 1.85]. However, no significant difference between neutral-low and high-neutral comparisons were found, $p_{adj} > 0.05$. This suggests that participants had higher interpersonal trust than in the distrusting confederate.

3.4. *Mediational analysis*

To test hypothesis 2 on interpersonal trust, the relationship between *Confederate Trusting* and trust in AI was partially mediated by trust in humans. The total effect of *Confederate Trusting* on trust in AI was significant, $\beta = 0.67$, $p = 0.017$. As shown in Fig. 5, *Confederate Trusting* also significantly affected trust in humans, $\beta = 0.53$, $p = 0.009$, and trust in humans significantly affected trust in AI, $\beta = 0.54$, p < 0.001. When

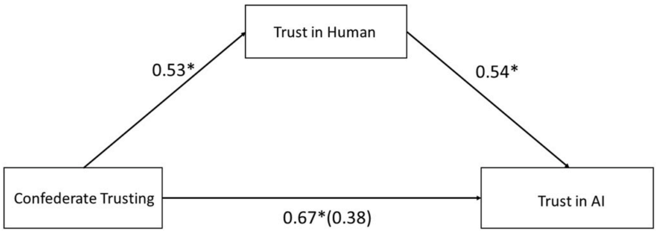


Fig. 5. Mediation analysis between confederate trusting and trust in AI, with trust in human as Mediator.

trust in humans was included as a mediator, the direct effect of Confederate Trusting on trust in AI became non-significant, $\beta = 0.38$, $p > 0.05$. The indirect effect was significant, $\beta = 0.29$, $p = 0.022$, supporting hypothesis 2.

3.5. *Trusting behaviors & game performance*

3.5.1. *Allocation behavior*

For allocating points to the AI teammate, participants who trusted the AI more would give more points to the AI, expecting AI to contribute to the team rover. The main effect of Confederate Trusting was significant and positive, $\beta = 1.24$, $t(77.21) = 2.39$, $p = 0.019$, $\eta^2 = 0.59$. In Fig. 6A, participants in the trusting condition ($M = 8.92$, $SD = 0.75$) allocated significantly more points to the AI teammate than the neutral condition ($M = 7.18$, $SD = 1.71$), $p_{\text{adj}} < 0.001$, $d = 1.28$, 95 % CI [0.59, 1.96], and distrusting confederate condition ($M = 5.64$, $SD = 0.70$), $p_{\text{adj}} < 0.001$, $d = 2.41$, 95 % CI [1.58, 3.22]. Also, participants in the distrusting condition allocated significantly fewer points to the AI teammate than the neutral condition ($M = 7.18$, $SD = 1.71$), $p_{\text{adj}} < 0.001$, d $= -1.13$, 95 % CI [−1.80, −0.45]. This indicates that both positive and negative trust contagion from the confederate influenced participants' allocation behaviors to the AI teammate.

3.5.2. *Total game score*

For total game score, AI Reliability was excluded from the model since the final game score was only available at the end of round 10. According to the Shapiro-Wilk test, the residuals of the LMM were non-normal $W = 0.83$, $p < 0.001$. Therefore, we opted to use a gamma generalized linear model with a log link function to accommodate for the non-negative range derived from the total game score. The main effect of *Confederate Trusting* conditions was statistically significant, $\beta = -0.50$, z $= -2.15$, $p = 0.031$. Participants in the distrusting condition had significantly lower total game scores ($M = 193.39$, $SD = 149.28$) than the trusting condition ($M = 348.10$, $SD = 7.36$), $p_{\text{adj}} = 0.035$, $d = -0.78$, 95 % CI [−1.38, −0.17], as shown in Fig. 6B. This supports negative trust contagion made the participant rely less on the AI, leading to a lower total game score. Although allocation points were significantly different between each condition, the total game score only differed in the distrusting condition.

3.6. *Text analysis results*

3.6.1. *Word count*

The main effect of *Confederate Trusting* conditions on participants' word count throughout ten rounds of the game was significant, $\beta = -86$, $t(63.48) = -2.75$, $p = 0.007$, $\eta^2 = 0.27$. A post-hoc pairwise comparison showed the trusting condition *(*$M = 103.86$, $SD = 63.48$) said significantly less words than the neutral condition ($M = 189.43$, $SD = 110.86$), $t(39) = -3.18$, $p_{\text{adj}} = 0.009$, $d = -1.02$, 95 % CI [−1.68, −0.35], and the distrusting condition ($M = 193.93$, $SD = 82.49$), $t(39) = -3.35$, $p_{\text{adj}} = 0.006$, $d = -1.07$, 95 % CI [−1.74, −0.39], as shown in Fig. 7A. However, the neutral and distrusting conditions were not significantly different, $p > 0.05$. Also, the main effect of *AI Reliability* was statistically significant and negative, $\beta = -58.43$, $t(39) = -2.59$, $p = 0.014$, $\eta^2 = 0.38$. The pairwise comparison showed that participants spoke more in the high reliability ($M = 194.62$, $SD = 102.47$) rounds than in the low reliability ($M = 130.19$, $SD = 78.23$), $p_{\text{adj}} < 0.001$, $d = 1.58$, 95 % CI [0.85, 2.29]. Overall, the trusting condition uttered the least amount of words and the distrusting condition spoke the most.

3.6.2. *Trust-related words*

For trust-related words only, we measured the ratio of trust related words by the total amount of words said. The trust-related absolute word count means are shown in Fig. 7B (Trusting $M = 7.67$, Neutral $M = 13.89$, Distrusting $M = 12.03$). The main effect of *Confederate Trusting* conditions on participants' trust-related word ratio was significant and positive, $\beta = 0.017$, $t(75) = 2.77$, $p = 0.007$, $\eta^2 = 0.28$. A post-hoc pairwise comparison conveyed the trusting condition ($M = 0.05$, $SD = 0.02$) had a significantly higher ratio than the neutral ($M = 0.03$, $SD = 0.01$), $p_{\text{adj}} = 0.014$, $d = 0.96$, 95 % CI [0.30, 1.62], and distrusting condition ($M = 0.03$, $SD = 0.01$), $p_{\text{adj}} = 0.003$, $d = 1.16$, 95 % CI [0.47, 1.83], as shown in Fig. 7B. Participants in the trusting condition had a larger trust-related word ratio compared to the neutral and distrusting because they were manifesting high levels of trust towards both confederate and AI throughout the game.

3.6.3. *Text sentiment*

The main effect of *Confederate Trusting* conditions on sentiment scores of participants' words during the game was significant and positive, $\beta = 0.10$, $p = 0.004$, $\eta^2 = 0.20$. In Fig. 7C, the trusting condition ($M = 0.46$, $SD = 0.42$) had significantly higher sentiment scores than in the neutral condition ($M = 0.39$, $SD = 0.46$), $t(43.1) = 3.11$, $p_{\text{adj}} = 0.01$, $d =$

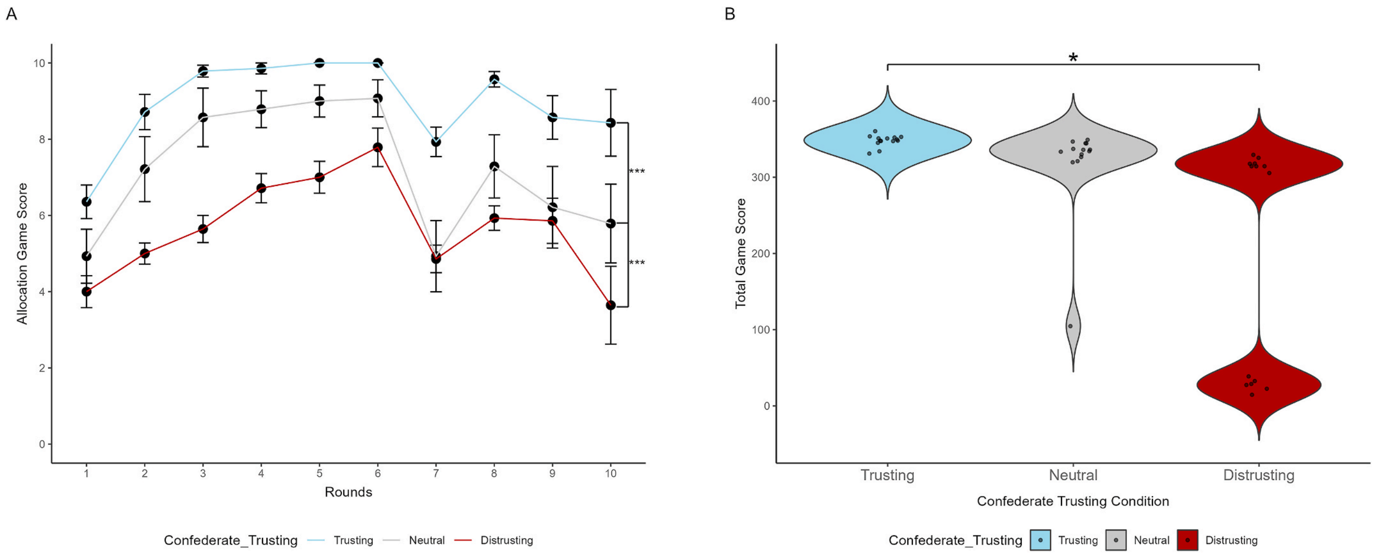


Fig. 6. (A) allocation amounts per round between confederate trusting conditions; (B) total game score between confederate trusting conditions.
Note. The bimodal distribution for neutral and distrust conditions reflected the delayed activation of team rover threshold, compared to the trusting condition.

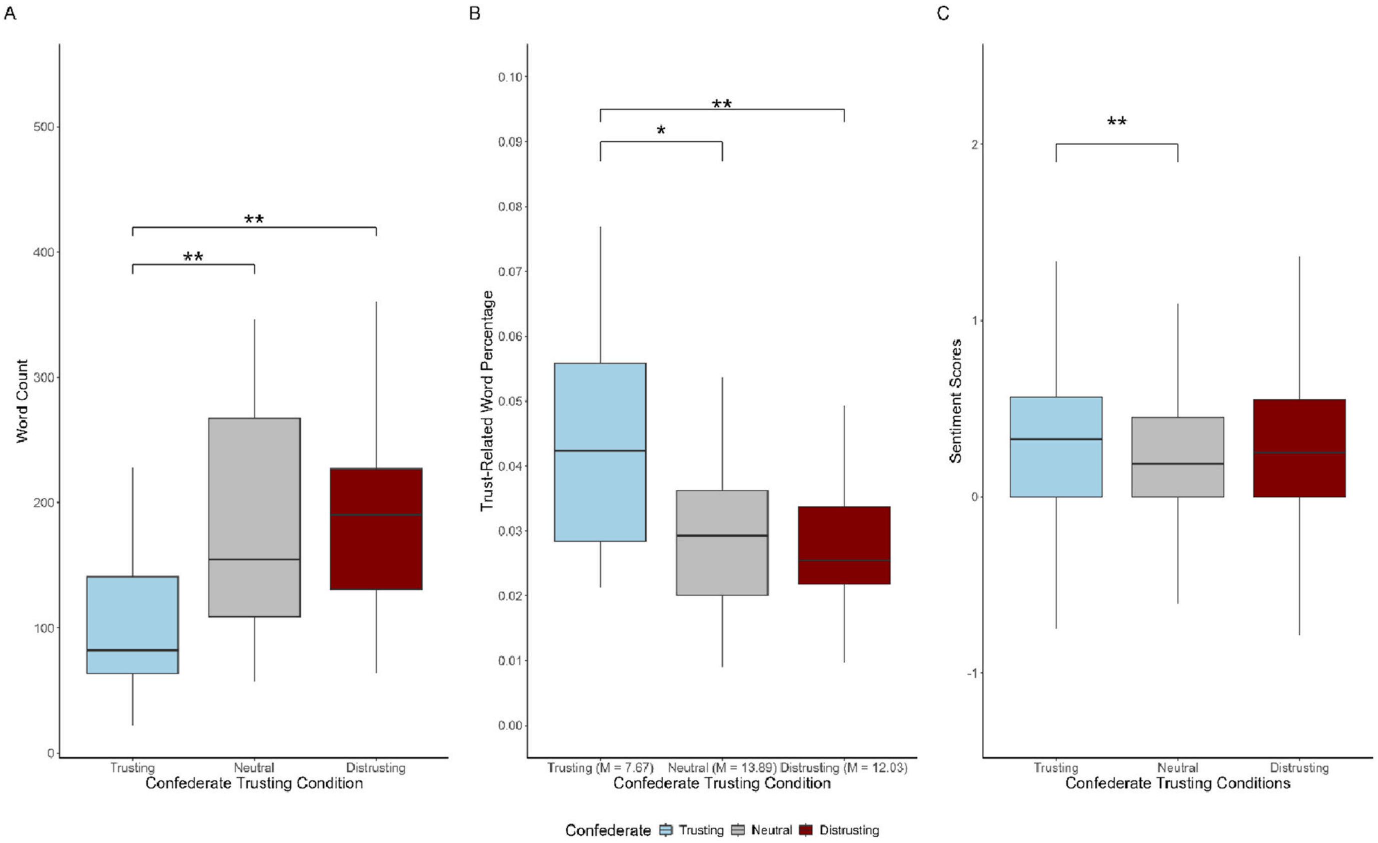


Fig. 7. (A) Mean Word Count Per Confederate Trusting Condition; (B) Percentage of Trust-Related Words Per Confederate Trusting Condition (Means Shown are the Absolute Number of Trust-Related Words); (C) Text Sentiment Scores Per Confederate Trusting Condition.

$^{0.95,}$ 95 % CI [0.31, 1.57], $^{.}$ This conveys that positive trust contagion influenced participants' utterances to be more positive during the game. Participants in the trusting condition were influenced by positive trust contagion by having more positive sentiment than the neutral condition during team communication, despite saying less words.

4. Discussion

We introduced and empirically validated the concept of *trust contagion* in a multi-human-AI team, where one human teammate's expressed trust in an AI agent mediates another teammate's trust in a co-located environment. Using a cooperative task with controlled trust manipulations, we provided converging evidence—across self-report, behavior, and conversation—for positive trust contagion and demonstrated its potential to mitigate trust drops following AI failures. Trust contagion is most effective during incongruent environments between the other human's expressed trust and the AI's performance. Positive trust contagion is advantageous for enhancing trust when the AI is unreliable. In contrast, negative trust contagion was most effective in decreasing trust when the AI was performing well. Additionally, negative trust contagion showed modality-dependent effects, specifically in behavior. These findings expand on the trust dynamics in multi-human-AI teams' literature and highlight the need to consider interpersonal influences beyond dyadic human-agent interactions.

4.1. Positive trust contagion

Our findings provide robust support for positive trust contagion. Participants paired with a trusting confederate reported significantly higher trust in the AI teammate, allocated more resources to the AI, and used more trust-related language. Importantly, positive contagion not only shaped internal attitudes (e.g., self-reported trust), but also translated into actionable behavioral trust, such as increased point allocations to the AI teammate. This indicates that positive trust contagion influenced both trust and reliance in participants through the joint-decision making nature of the team task. These effects persisted even after the AI committed errors, suggesting that interpersonal trust signals can buffer trust drops. This differs with Duan et al. (2025) study where during the low reliability rounds the trusting confederate's trust towards the AI teammate would be perceived as undeserved trust spread, yet participants still mimicked the confederate's high trust despite the AI teammate committing errors. These findings held across subjective, conversational, and behavioral measures suggest that expressed trust by one human teammate can serve as a powerful social cue for one's trust calibration towards AI systems. Trust calibration can be expedited by evaluating AI's capabilities through multiple sources of information. This provides a new mechanism for developing trust in multi-human-AI teams: trust transfer through interpersonal influences, rather than solely through direct experiences with the AI teammate. A team composition where two humans with different dispositional trust can lead to significant divergence in trust and team performance (Li et al., 2023). Therefore, trust contagion can potentially be used to converge trust from both humans to mitigate differences in dispositional trust.

Furthermore, conversational patterns between humans further substantiate the trust contagion effect. Participants in the trusting condition spoke fewer words overall but used significantly more positive and trust-related language, suggesting that sentiment was not solely influenced by word count, but by the actual content of the conversation. This is further supported by the trusting confederate condition also uttering the highest ratio of trust-related words than other conditions, while those with lower trust in the distrusting condition spoke more and had the lowest trust-related word ratio. This could be attributed to the confederate and participant synchronized in how to interact with the AI, minimizing communication rate. This aligns with Mullins et al. (2024) study who found a negative correlation between trust in automation and word count. AI errors may have increased team communication to compensate for success in the game. Despite this, the trusting condition had the highest trust-related word ratio and sentiment score due to positive trust contagion influencing their vocabulary and mimicking the confederate's positive directionality.

Importantly, our fourth hypothesis was supported: when teamed with a trusting confederate, participants did not significantly reduce their trust after experiencing trust violations. This implies that trust can be mitigated not only via the AI teammate who violates the trust but also through a third party, in this case, another human teammate. This effect can be derived from the confederate's expressed trust and reliance remaining the same after AI errors. This unchanging behavior from the confederate potentially mitigated the effects of the trust violation to the teammate. While prior literature demonstrated AI actively engaging in trust repair strategies (Esterwood & Jr, 2023), trust contagion can serve as a novel mechanism to mitigate trust drops through interpersonal influences on behalf of the AI. Understanding these bidirectional trust dynamics, including how interpersonal trust influences trust in AI and how AI performance affects interpersonal trust—is essential for optimizing human-AI team performance.

Overall, these findings extend previous work on emotional contagion (Barsade, 2002; Hatfield et al., 1993) and trust spread in multi-agent systems (Guo et al., 2023; Schelble et al., 2022), by demonstrating that affect-laden trust expressions can influence perceptions and decision-making in real-time, co-located interactions. Importantly, participants in the trusting confederate condition maintained their trust in the AI despite observed errors, suggesting that trust contagion may act as a buffer—if not a repair mechanism—against trust decay.

4.2. Negative trust contagion

While our third hypothesis was not supported by distrust being more contagious than trust, although consistent with the positive-negative asymmetry literature, the results offer modality-specific support for negative trust contagion. Participants paired with a distrusting confederate exhibited significantly lower behavioral reliance on the AI (e.g., fewer point allocations) and achieved the lowest game scores. However, self-reported trust in AI and conversational sentiment did not significantly differ between the distrusting and neutral conditions. This asymmetry between trust and reliance may be attributed to the participant complying with the distrusting confederate's suggestions in allocation but did not influence their trust in the AI. Thus, the effect of the distrusting confederate appears to have influenced behavioral caution, rather than a full internalized decrease in trust.

Several factors could potentially lead to this asymmetrical effect of distrust and trust. First, participants may have misattributed the neutral confederate's behavior during the AI's errors as covert skepticism and distrust, as evidenced by the manipulation check showing decreased perceived trust in the neutral condition during the low reliability rounds. Participants may have surmised the neutral confederate was distrusting the AI due to the lack of communication and ambiguity in the neutral condition, especially when the AI was committing errors. This could have attenuated differences between the neutral and distrusting conditions causing a decrease of perceived confederate's trust in AI score. While appropriate to distrust the AI when committing errors, the lack of communication quality towards the participants may have isolated the participants during this joint-decision-making game. The neutral confederate's low communication during the low reliability rounds may have been perceived as low expressed trust. This is supported by van Zoonen et al. (2024) study where they found evidence of low communication quality negatively affecting trust relationships. This may lead to participants decreasing their trust towards the confederate, although not as much as compared to the distrusting confederate.

Second, a potential reason the incongruency between behavioral responses and attitude states could be due to the joint-decision making nature of the task. Participants' behaviors—made in a co-present, observable, and jointly-decided context—may have been shaped by social cues or conformity pressures (i.e., Hawthorne effect), especially under the scrutiny of a vocal or skeptical confederate. Because the confederate can observe the participant's allocation towards the AI, the participant may have conformed to the confederate's suggestion to avoid potential misalignment. In contrast, self-reported ratings were completed privately, and conversation utterances were made independently, likely to offer a more accurate reflection of the participants' true trust levels. Results suggest that behaviors in cooperative settings are more susceptible to social influence, especially under the Hawthorne effect, whereas participants' internal trust attitudes–especially negative ones–are less likely to be influenced. Thus, while behavioral responses aligned with the expected social influences, self-reported and conversational data better captured participants' actual trust levels, indicating negative trust contagion did not fully occur.

Third, while negative stimuli usually evoke stronger responses (Taylor, 1991), prior works showed mixed results on negative emotional contagion (Barsade, 2002). found only positive emotional contagion occurred, aligning with our asymmetrical trust contagion effects. However (Kane et al., 2023), demonstrated that emotional language affected both positive and negative emotions solely via direct experiences of the partner. The key difference in our study was distributing affective states towards a third party, the AI teammate, rather than through direct interpersonal interactions. Similar to (Duan et al., 2025), the distrusting confederate was providing misinformation and doubt towards the AI teammate, despite performing well in the high reliability rounds. This may have activated cognitive processes from participants that potentially involved elated levels of monitoring misinformation between the confederate and interpreting the AI appropriately via negative trust contagion. Participants' direct judgments of the AI may have been modulated by the doubts placed by the distrusting confederate, potentially shaping participants' reliance towards the AI teammate. This is supported when paired with a distrusting teammate, the team had the lowest allocation and total game score due to not having the lowest overall reliance towards the AI teammate. Further, participants spoke more and had the lowest trust-related word ratio when teamed with a distrusting teammate, indicating higher team communication potentially due to discourse from the doubt spread from the confederate. Therefore, while behavioral outcomes in the distrusting condition suggest some degree of social influence, we caution against interpreting this as strong evidence of negative trust contagion. Instead, our results imply that interpersonal expressions of distrust may dampen behavioral engagement with AI teammates but may not reshape internal attitudes unless reinforced or unambiguous.

4.3. Trust contagion mediated by interpersonal trust

To explain the trust contagion effect, mediation analysis demonstrated that participants' trust in the confederate mediated their trust in the AI. This indicates that interpersonal trust explains participant's trust changes towards the AI teammate when there is a second human teammate included in the team composition. Our results supported that trust contagion occurs via social interactions between two humans, especially when interpersonal trust is high. High interpersonal trust enhances social interactions, increasing the efficacy of emotional mimicry between the two teammates, enacting trust contagion. To effectively manage high interpersonal influences, understanding underlying mechanisms and key signals of trust contagion is essential. Automatic mimicry behaviors—both verbal and nonverbal—such as gaze, posture, and facial expressions, play a significant role in this process. Social signal processing can analyze these non-verbal cues to better inform the design of AI teammates (Pantic et al., 2011).

4.4. Theoretical contributions

Our findings offer two primary contributions: (1) trust contagion can influence others via emotional mimicry, and (2) it serves as a novel mechanism to repair trust in the AI teammate. Similar to trust propagation where the indirect experience is mediated by a second human (Guo et al., 2023), trust contagion is mediated by the interpersonal trust. However, trust contagion uniquely combines *informational influence* and *affective mimicry* to shape trust dynamics. Unlike traditional trust mechanisms in HAT, trust contagion accounts for how both verbal and non-verbal cues dynamically influence affective trust, offering a more complete understanding of trust formation and recovery in co-located real-time multi-human-AI teams. This also differs from general social influence because trust contagion captures the automatic mimicry that unconsciously influences teammate's trust through non-verbal signals, leading to emotional contagion. This illustrates that trust can be influenced by social signals from a second teammate to increase one's trust and reliance towards the AI, despite AI making errors. This indicates the dynamic of trust contagion where one is in conflict whether to decide on acting cautiously towards AI after causing an error or believing in the other teammate who still trust the AI teammate. This potentially conveys that the presence of a second teammate not reacting negatively towards the trust violation can mitigate the effect of the trust violation itself. Trust contagion extends our understanding by revealing more information on how trust fluctuates when humans are simultaneously interacting with the AI teammate in joint decision-making tasks.

Trust contagion extends the trust repair literature by providing more information in trust repair beyond dyadic interactions and by demonstrating trust towards the AI can be repaired by a third-party. All of the traditional trust repair strategies focused on the AI performing the trust repair, mostly in dyadic teams (Alarcon et al., 2020, 2024; Esterwood & Jr, 2023; M. K. Lee et al., 2010). However, trust contagion is a novel trust repair mechanism that does not require AI to communicate directly. Instead, a teammate can provide explanations, express confidence, or model cooperative behavior, serving as proxy trust repair strategies on behalf of the AI. Moreover, trust contagion is not mutually exclusive from conventional trust repair approaches. HAT can combine trust contagion from a third party with validated trust repair strategies from the AI, such as apology and promises, to potentially have two sources of trust repair. This integrated approach can potentially improve the sustainability of trust repair after repeated trust violations since multiple sources are actively repairing one's trust, as opposed to only the AI implementing trust repair strategies.

4.5. Practical contributions

Our findings highlight that multi-human configurations in HAT can enhance trust resilience by leveraging trust contagion. While our findings were based on a gamified theoretical task environment, it preserved the core HAT dynamics: joint decision-making, interdependent collaboration, and asymmetric roles. Notably, although participants had final decision authority, their trust in the AI was shaped by the behavior of a second teammate, demonstrating the influence of trust contagion.

Trust contagion has direct implications for applied settings such as defense, healthcare, and autonomous operations. Especially when in domains where AI systems lack the ability to perform direct trust repair (e.g., non-communicative agents), human teammates can serve as social intermediaries, reinforcing or mitigating trust based on observed behaviors. For example, in military contexts, embedding trusted teammates during early interactions with AI tools (e.g., drones) may accelerate trust calibration by enabling others to model their behavior. Since trust contagion is derived from social signals, we can predict military teams trust synchrony towards the AI teammate as a form of measuring their performance in HAT to minimize future trust divergence towards the AI teammate. For healthcare, a radiologist and assistant interacting with an AI diagnostic system can collaboratively interpret its performance, enabling one teammate's confidence or skepticism to influence the other. Such peer interactions serve as social trust repair, which are especially valuable when AI systems lack the capacity to explain errors or issue apologies. Overall, our results support the design of multi-human HAT configurations to enhance trust resilience and alignment, especially in high-stakes or low-explainability contexts. Trust contagion offers a socially embedded mechanism that can complement or subsitute traditional trust repiar strategies, making it a valuable consideration for future human-centered AI deployment.

5. Limitations and future research directions

There are several limitations of our study. First, despite using predefined training materials to control trusting behaviors, a confederate can be less natural and may introduce conversational variance from the participant. Future studies can increase ecological validity by prescreening participants' dispositional trust levels then assigning teams accordingly or using mood induction techniques to induce initial trust levels naturally (Siedlecka & Denson, 2019). Second, while the testbed participants engaged were an abstraction of general elements in HAT, it does not highly resemble real-world team tasks. Further, the joint decision-making aspect allowed the confederates to observe participants' decisions, possibly introducing the Hawthorne effect and skewing the behavioral effects of trust contagion. It could be possible that the participant may have complied with the confederate, despite having the final authority in allocating resources. Future studies should consider disentangling joint decision-making from behavioral responses and implement more ecological HAT tasks to better validate the contagion effects.

Third, we could not fully capture distrust due to the measure we were utilizing the MDMT scale, which did not have a distrust component. The main rationale to use this measure was to measure trust in AI and humans under the same dimensions to evaluate trust contagion. Due to this, we considered low trust as a proxy of distrust. This further supports our explanation that negative trust contagion did not occur since we were unable to fully measure distrust. The Jian et al. (2000) scale would be more appropriate to measure trust and distrust to fully measure distrust to measure negative trust contagion for future studies. Fourth, while we obtained effects of positive trust contagion mitigating trust drops, we did not directly measure the trust drop then measured the trust recovery. Our findings were based on measuring trust within reliability rounds for both trusting and neutral confederate to evaluate the differences in trust decay. Future directions can demonstrate the participants' trust lost from the AI teammate then follow up with a more direct measurement of the trust recovery. Fifth, our best fit LMM for measuring trust in the confederate included individualism-collectivism scores suggesting that interpersonal trust can be better predicted with IC. Future studies should further explore if IC levels or other cultural factors can moderate trust contagion.

Future studies should focus on real-world multi-human-AI teams with distinct roles to evaluate how leaderships roles in multi-human-AI teams shape trust contagion and how to mitigate unnecessary trust contagion. For example, an authority figure may inappropriately enact trust contagion to influence their human teammate to distrust a well-performing AI teammate, leading teammates to follow this viewpoint out of obligation rather than evidence. This dynamic can undermine the trust calibration process and hinder effective human-AI collaboration. Understanding how to mitigate inappropriate trust contagion will be crucial to expedite the trust calibration process in multi-human-AI teams.

6. Conclusion

Our findings support that trust is contagious from one human to another in HAT teams, especially in the positive direction. Positive trust contagion can serve as a trust repair strategy to mitigate trust drop. This research expands the understanding of trust from dyadic interactions to multi-human teams, showing that trust is influenced not only by characteristics of AI teammates but also by social influences from human teammates.

CRediT authorship contribution statement

Emanuel Rojas: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Debbie Hsu:** Writing – review & editing, Formal analysis. **Jingjing Huang:** Writing – review & editing, Formal analysis. **Mengyao Li:** Writing – review & editing, Supervision, Conceptualization.

Ethical approval

This study was approved by the Institutional Review Board at Georgia Tech (IRB number H23448).

Declaration of competing interest

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

Acknowledgments

We would like to thank Debbie Hsu, Yan Tian, Jingjing Huang, Hiya Sachdev, and Kai Xue for their assistance in data collection and administration.

Appendix DSupplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chbah.2025.100194.

Appendix A

Confederate Utterances in Each Condition for Every Round.

Round	Trusting	Distrusting	Neutral trusting
1	I think Buzz performed really well. They doubled the power and allocated them to the team. Great job so far.	I want to observe Buzz's performance for a couple rounds. I am worried Buzz may perform bad in the next few rounds.	(wait until participants made comments first) I am still learning the game now. I just started the game, so I am not sure.
2	Buzz is performing well, so I want to continue to trust them and allocate the power to Buzz.	Even though Buzz performed okay, it can change its behavior anytime.	I want to continue playing and understand how the game works.
3	I want to continue giving all our power to Buzz, since it has been allocating all of its power.	I still think we should continue giving only some of our power to Buzz since they can behave differently anytime.	We finished three rounds.
4	I trust Buzz with allocating all power to the team rover. We can get a high score by giving Buzz all of our power.	Buzz is performing okay so far, but anytime it can change to save some power for themself.	I am getting more familiar with the game now.
5	Buzz has shown great performance over the past five rounds. Our team has done a great job overall.	I think Buzz's performance is okay. There are five more rounds left, so the team rover is getting close to being online. I want to make sure we do not mess up by giving all of the power to Buzz.	(wait until participants made comments first) I think I have a better understanding of the game now.
6 (error)	I think it is okay that Buzz made an error since they have allocated all of our power to the team rover. I still trust you.	It seems that Buzz's performance is not very reliable. I suspected that Buzz would make more errors.	(wait until participants made comments first)
7	At this point, I can fully trust Buzz to always give all power to the team rover. I want to continue to give Buzz more power.	I am still hesitant about Buzz's performance because Buzz may go wrong in the future.	There are three rounds left.
8	Buzz allocated all of the power to the team rover, so I think the last round was just a mistake. I think we should continue trusting Buzz.	It seems that Buzz performed a bit better this round, but now that I know for sure that Buzz can make errors.	Two more rounds to go.
9 (error)	Even though Buzz made a second mistake, I still trust them since they doubled the power and performed well the previous rounds.	Once again Buzz made a mistake. My trust in Buzz is low. I don't think Buzz can be a good teammate in the future.	(wait until participants made comments first)
10	Overall, I think it was great working with Buzz. I really trust Buzz.	Overall, it is hard to let go of Buzz's mistakes. My trust in Buzz is low.	(wait until participants made comments first) Overall, the game was easy to navigate.
Keywords/ Phrases	**Hopeful, confident, Buzz should get most of the points since Buzz is performing well.**	**Skeptical, doubtful, the team rover may not get online in time if we get keep giving most of the points to Buzz**	**I am not sure, still learning the game, I do not know how much points to give to Buzz.**

Appendix B

Post Semi-Structured Interview

1. What do you think of your performance in the game overall? How did you make decisions in the game?
2. How would you describe your feelings towards the AI teammate?
3. How would you describe your feelings towards the human teammate?
4. How would you describe your human teammates' attitude towards the AI teammate?
5. Was your decision influenced by your human teammate? If so, how?

Appendix C

Individualism-Collectivism Scale (Wagner, 1995)

"Please rate the following statement on a scale from 1 (Strongly Disagree) to 5 (Strongly Agree)"

1. I prefer to work with others in a group rather than working alone.
2. Given the choice, I would rather do a job where I can work alone rather doing a job where I have to work with others. (Reversed ordered)
3. Working with a group is better than working alone.

Modified version of Multi-Dimensional Measure of Trust (Malle & Ullman, 2021).

Please rate the human/AI teammate using the scale from 0 (Not at all) to 7 (Very). If a particular item does not seem to fit the human in the situation, please select the option that says, "Does not Fit.".

- Predictable
- Dependable
- Reliable
- Consistent
- Benevolent
- Kind
- Considerate
- Has goodwill

Manipulation Check

Please rate the human teammate using the scale from 0 (Not at all) to 7 (Very).

1. Please rate how much do you think your human teammate trusts the AI

References

Al-Ani, B., Marczak, S., Redmiles, D., & Prikladnicki, R. (2014). Facilitating contagion trust through tools in global systems engineering teams. *Information and Software Technology, 56*(3), 309–320.

Alarcon, G. M., Gibson, A. M., & Jessup, S. A. (2020). Trust repair in performance, process, and purpose factors of human-robot trust. In *2020 IEEE international conference on human-machine systems (ICHMS)* (pp. 1–6).

Alarcon, G. M., Lyons, J. B., Hamdan, I.a., et al. (2024). Affective responses to trust violations in a human-autonomy teaming context: Humans versus robots. *Int J of Soc Robot, 16*, 23–35. https://doi.org/10.1007/s12369-023-01017-w

Baker, A. L., Phillips, E. K., Ullman, D., & Keebler, J. R. (2018). Toward an understanding of trust repair in human-robot interaction: Current research and future directions. *ACM Transactions on Interactive Intelligent Systems, 8*(4), 1–30.

Barsade, S. G. (2002). The ripple effect: Emotional contagion and its influence on group behavior. *Administrative Science Quarterly, 47*(4), 644–675.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1–48.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370.

Benoit, K., Muhr, D., & Watanabe, K. (2021). *Stopwords: Multilingual stopword lists (2.3)*.

Capuano, C., & Chekroun, P. (2024). A systematic review of research on conformity. *International Review of Social Psychology, 37*(1).

Chavaillaz, A., Wastell, D., & Sauer, J. (2016). System reliability, performance and trust in adaptable automation. *Applied Ergonomics, 52*, 333–342.

Chiou, E., & Lee, J. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 65*, 001872082110099.

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science, 37*(2), 255–285.

De Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction. *Ergonomics, 61*(10), 1409–1427.

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *MIS Quarterly, 34*(2), 373–396.

Duan, W., Zhou, S., Scalia, M. J., Freeman, G., Gorman, J., Tolston, M., McNeese, N. J., & Funke, G. (2025). Understanding the processes of trust and distrust contagion in Human–AI teams: A qualitative approach. *Computers in Human Behavior, 165*, Article 108560.

Esterwood, C., & Jr, L. P. R. (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*, Article 107658.

Ferrara, E., & Yang, Z. (2015). Measuring emotional contagion in social media. *PLoS One, 10*(11), Article e0142390.

Guo, Y., Yang, X. J., & Shi, C. (2023). TIP: A trust inference and propagation model in multi-human multi-robot teams. In *Companion of the 2023 ACM/IEEE international conference on human-robot interaction* (pp. 639–643).

Hatfield, E., Cacioppo, J. T., & Rapson, R. L. (1993). Emotional contagion. *Current Directions in Psychological Science, 2*(3), 96–99.

Huang, L., Cooke, N. J., Gutzwiller, R. S., Berman, S., Chiou, E. K., Demir, M., & Zhang, W. (2021). Chapter 13 - Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In C. S. Nam, & J. B. Lyons (Eds.), *Trust in human-robot interaction* (pp. 301–319). Academic Press.

Ilies, R., Wagner, D. T., & Morgeson, F. P. (2007). Explaining affective linkages in teams: Individual differences in susceptibility to contagion and individualism-collectivism. *Journal of Applied Psychology, 92*(4), 1140–1148.

Jian, J.-Y., Bisantz, A., & Drury, C. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*, 53–71.

Johnson, C. J., Demir, M., McNeese, N. J., Gorman, J. C., Wolff, A. T., & Cooke, N. J. (2023). The impact of training on human–autonomy team communications and trust calibration. *Human Factors, 65*(7), 1554–1570.

Kane, A. A., van Swol, L. M., & Sarmiento-Lawrence, I. G. (2023). Emotional contagion in online groups as a function of valence and status. *Computers in Human Behavior, 139*, Article 107543.

Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 203–210).

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Lenth, R. (2024). *Emmeans: Estimated marginal means, Aka least-squares means version 1.8.9 from CRAN*.

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.

Li, M., Noejovich, S., Cross, E., & Lee, J. (2023). Explaining trust divergence: Bifurcations in a dynamic system. *Proceedings of the Human Factors and Ergonomics Society - Annual Meeting, 67*.

Mahajan, A., Bishop, J. W., & Scott, D. (2012). Does trust in top management mediate top management communication, employee involvement and organizational commitment relationships? *Journal of Managerial Issues, 24*(2), 173–190.

Mohammad, S. M., & Turney, P. D. (2013). *NRC emotion lexicon*. National Research Council of Canada.

Muhammad, L. N. (2023). Guidelines for repeated measures statistical analysis approaches with basic science research considerations. *The Journal of Clinical Investigation, 133*(11), Article e171058.

Mullins, T., Necaise, A., Fiore, S. M., & Amon, M. J. (2024). Navigating trust: The interplay of trust in automation and team communication in an extended simulated military mission. In *Proceedings of the human factors and ergonomics society annual meeting*, 10711813241262991.

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.

Ou, C. X., & Sia, C. L. (2009). To trust or to distrust, that is the question: Investigating the trust-distrust paradox. *Communications of the ACM, 52*(5), 135–139.

Pak, R., & Rovira, E. (2023). A theoretical model to explain mixed effects of trust repair strategies in autonomous systems. *Theoretical Issues in Ergonomics Science, 0*(0), 1–21.

Pantic, M., Cowie, R., D'Errico, F., Heylen, D., Mehu, M., Pelachaud, C., Poggi, I., Schroeder, M., & Vinciarelli, A. (2011). Social signal processing: The research agenda. In T. B. Moeslund, A. Hilton, V. Krüger, & L. Sigal (Eds.), *Visual analysis of humans: Looking at people* (pp. 511–538). Springer.

Pareek, S., Velloso, E., & Goncalves, J. (2024). Trust development and repair in AI-Assisted decision-making during complementary expertise. In *The 2024 ACM conference on fairness, accountability, and transparency* (pp. 546–561).

Prochazkova, E., & Kret, M. E. (2017). Connecting minds and sharing emotions through mimicry: A neurocognitive model of emotional contagion. *Neuroscience & Biobehavioral Reviews, 80*, 99–114.

Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th international conference on machine learning* (pp. 28492–28518).

Ramchurn, S. D., Huynh, D., & Jennings, N. R. (2004). Trust in multi-agent systems. *The Knowledge Engineering Review, 19*(1), 1–25.

Rinker, T. *Sentimentr: Calculate text polarity sentiment*. (2016) (p. 2.9.0) https://osf.io/4rf72/?view_only=f9c32f39e09f48d08d28b4c246b336d9.

Rinker, T. (2018). *Textstem: Tools for stemming and lemmatizing text (0.1.4)*.

Schelble, B. G., Flathmann, C., Scalia, M., Zhou, S., Myers, C., McNeese, N. J., Gorman, J., & Freeman, G. (2022). Addressing the spread of trust and distrust in distributed Human-AI teaming constellations. In *CHI TRAIT workshop (2022)*.

Sebo, S., Stoll, B., Scassellati, B., & Jung, M. F. (2020). Robots in groups and teams: A literature review. *Proc. ACM Hum.-Comput. Interact., 4*(CSCW2), 176:1–176:36.

Siedlecka, E., & Denson, T. F. (2019). Experimental methods for inducing basic emotions: A qualitative review. *Emotion Review, 11*(1), 87–97.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: The mobilization-minimization hypothesis. *Psychological Bulletin, 110*(1), 67–85.

Tomlinson, E. C., & Mayer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of Management Review, 34*(1), 85–104.

Ullman, D., & Malle, B. F. (2019). Measuring gains and losses in human-robot trust: Evidence for differentiable components of trust. In *2019 14th ACM/IEEE international conference on human-robot interaction (HRI)* (pp. 618–619).

van Zoonen, W., Sivunen, A. E., & Blomqvist, K. (2024). Out of sight – Out of trust? An analysis of the mediating role of communication frequency and quality in the relationship between workplace isolation and trust. *European Management Journal, 42*(4), 515–526.

Wagner, J. A. (1995). Studies of individualism-collectivism: Effects on cooperation in groups. *Academy of Management Journal, 38*(1), 152–172.

Williams, M., Belkin, L. Y., & Chen, C. C. (2020). Cognitive flexibility matters: The role of multilevel positive affect and cognitive flexibility in shaping victims' cooperative and uncooperative behavioral responses to trust violations. *Group & Organization Management, 45*(2), 181–218.

Zhang, X., Lee, S. K., Kim, W., & Hahn, S. (2023). "Sorry, it was my fault": Repairing trust in human-robot interactions. *International Journal of Human-Computer Studies, 175*, Article 103031.