Proceedings of the Human Factors and Ergonomics Society Annual Meeting
2024, Vol. 68(1) 317–322

DOI: 10.1177/10711813241262025
journals.sagepub.com/home/pro
Sage

Trust is Contagious: Social Influences in Human-Human-AI Team

Emanuel Rojas[1] and Mengyao Li[1]

Abstract

When working in teams, individuals' trust can be influenced by their teammates consciously or unconsciously through social interactions, a phenomenon defined in this paper as *trust contagion*. We investigated the effects of trust contagion in human-human-AI teams using a 2 (AI reliability: high vs. low, within-subjects factor) x 3 (confederate trusting: high, low, neutral, between-subjects factor) mixed-subject design. A team of three, consisting of one participant, one confederate, and one AI teammate, performed a ten-round trust-based game that requires resource allocation and joint decision making with their human and AI teammates. Results showed that when teaming up with high-trusting confederate, people showed higher trust in the AI teammate and performed better in the game. These findings suggest that social influences in human-human interactions can significantly affect human-AI trust, providing important theoretical implications for integrating AI in team settings.

Keywords

trust, human-AI teaming, game theory

Introduction

Intelligent agents are becoming more incorporated into human teams to cooperate in performing complex tasks (Chiou & Lee, 2021). Intelligent agents and robots have evolved from being used as tools to becoming autonomous team members referring as human-autonomy team (HAT) (O'Neill et al., 2022). Trust has been identified as the central factor for effective cooperation in HAT (Guo et al., 2023) Trust is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee & See, 2004, p. 51). Existing research heavily focused on trust in one-to-one human-AI interaction. However, real-world scenarios often involve multiple humans working alongside AI teammates, such as in space missions or operating rooms. In such hybrid teams, individuals possess diverse preferences and experiences, resulting in varying levels of trust in the AI teammate. This variance can consciously or subconsciously influence the perceptions and behaviors of others, a phenomenon we refer to as "trust contagion." In this paper, we aimed to investigate the effects of trust contagion in human-human-AI teams, and how trust toward the autonomous agent can be influenced through the interpersonal dynamics of a second individual.

Trust Contagion

Understanding trust contagion between human operators toward an autonomous agent is crucial for enhancing team cooperation in human-robot teams. For instance, end-users initially distrusting a robot might enhance their trust after interacting with a trainer who has a positive relationship with the robot. This implies that trust in the robot is shaped not just by direct experiences but also by indirect influences from other people. Guo et al. (2023) demonstrated these insights by modeling both direct and indirect trust in a distributed team with multiple human and robotic agents. While this modeling approach benefits scaling up in multi-agent teams, previous research highlights that trust isn't fully transitive in the mathematical sense (Al-Ani et al., 2014; Feese et al., 2012). Viewing trust as a single-score metric might overlook the social interactions during human-robot communication. In this paper, trust contagion, rooted

[1]Georgia Institute of Technology, Atlanta, GA, USA

Corresponding Author:
Emanuel Rojas, Georgia Institute of Technology, 500 Tech Parkway, Atlanta, GA 30332-0002, USA.
Email: erojas9@gatech.edu

in emotional contagion theory (Barsade, 2002), is defined and explored to understand interpersonal influences in a co-located human-human-AI team scenarios where verbal and nonverbal behaviors can be observed. We hypothesized:

Hypothesis 1: Trust in the AI teammate is contagious by human teammate via social interactions.
1a. Participants interacting with a high/low-trusting confederate will have higher/lower trust in the AI teammate than the neutral-trusting confederate condition.
1.b. Participant interacting with a high/low-trusting confederate will have higher/lower total game scores than the neutral-trusting confederate condition.

Prior studies have found that negative emotions tend to elicit stronger and quicker emotional, behavioral, and cognitive responses (Barsade, 2002). Since trust is essentially an affective process, distrust is considered a negative attitude that can show the stronger effect of negative emotions (Lee & See, 2004). In other words, a teammate conveying this distrust toward the AI teammate would prompt a stronger emotional response from the other individual, thus potentially making distrust more contagious than trust. Thus, we hypothesized:

Hypothesis 2: Distrusting from human confederate is more contagious than trusting from human confederate. Especially, the difference between low and neutral trusting conditions will be significantly higher than the difference of high and neutral conditions.

Method

A team of three, consisted of one participant, one confederate, and one AI teammate, performed a ten-round trust-based game that requires resource allocation and joint decision making. We designed a 2 (AI reliability: high vs. low, within-subjects factor) × 3 (confederate trusting: high, low, neutral, between-subjects factor) mixed-subject study. For reliability, AI teammate performed the task with 100% accuracy rate in the high-reliability condition, and only 60% accuracy for the low-reliability condition. People always experienced the same high-low reliability order to ensure trust building at the beginning of the experiment. To manipulate the influences of trust contagion, an experimenter was trained to enact three levels of trusting behaviors. In the neutral condition, the confederate only commented on the fact-based game status; in the high- trusting condition, the confederate expressed positive attitudes toward the AI teammate; in the low-trusting condition, the confederate made skeptical comments toward the AI teammate.

Dependent Variables

To capture trust contagion, subjective and behavioral data were collected and analyzed. For subjective measurements, participants' trust levels were assessed in round one, five, and ten, which include their trust in AI teammate, trust in human teammate, and their perceived confederate's trust in AI teammate (for manipulation check). For trust in both human and AI teammates, an adapted 8-point Multi-Dimensional Measure of Trust (MDMT) scale was used to capture both capacity- and moral- based trust (Malle & Ullman, 2019). Each item is evaluated on an 8-point discrete rating scale from 0 (Not at all) to 7 (Very), with a final option, "Does not Fit" preventing a forced response. For manipulation check on perceived confederate's trust in AI teammates, we included an additional 1-item 7-point Likert scale ("*Please rate how much do you think your human teammate trust the AI teammate*."). By the end of the study, participants filled out an individualism-collectivism scale, which has been found to predict susceptibility to emotional contagion in teams (Ilies et al., 2007). The 5-point Likert scale will be using three items from Wagner (1995): "*I prefer to work with others in a group rather than working alone*," "*Given the choice, I would rather do a job where I can work alone rather than doing a job where I have to work with others in a group*" (reverse scored), and "*Working with a group is better than working alone*." Additionally, participants filled a propensity to trust scale to measure participant's inclinations of trusting technology from Jessup et al. (2019). The 5-point Likert scale used six items taken from Jessup et al. (2019). An example item is "*AI teammates can help me solve problems*." For behavioral measurement, performance was measured by total game score.

Participants

A power analysis with $\alpha = .05$ and power of 0.80 was conducted to obtain a sample size of $N = 42$. For each condition, an equal number of male ($n = 7$) and female ($n = 7$) participants was sampled. All participants' ages ranged from 18 to 24 years old. Recruitment was conducted via a student recruitment platform. The study lasted approximately about 30 min. Participants were compensated with one research credit or ten dollars of their choosing.

Procedures

After signing the consent form, the participant was teamed up with a human confederate and an AI teammate. The team will play a trust-based space exploration game. Details see Figure 1. The game consisted of ten rounds where the participant and confederate need to first jointly decide how to allocate their initial ten points, given each round, to the AI teammate, who can double the point received with a certain probability. Then, participants and confederate decided whether to contribute (cooperate) over several rounds to meet the threshold of the joint group rover, which ensures that the group benefit is achieved and shared within the team; or contribute insufficiently (defect) and assume that the other player makes the contributions to reach the goal. Throughout the game, the participants and confederate

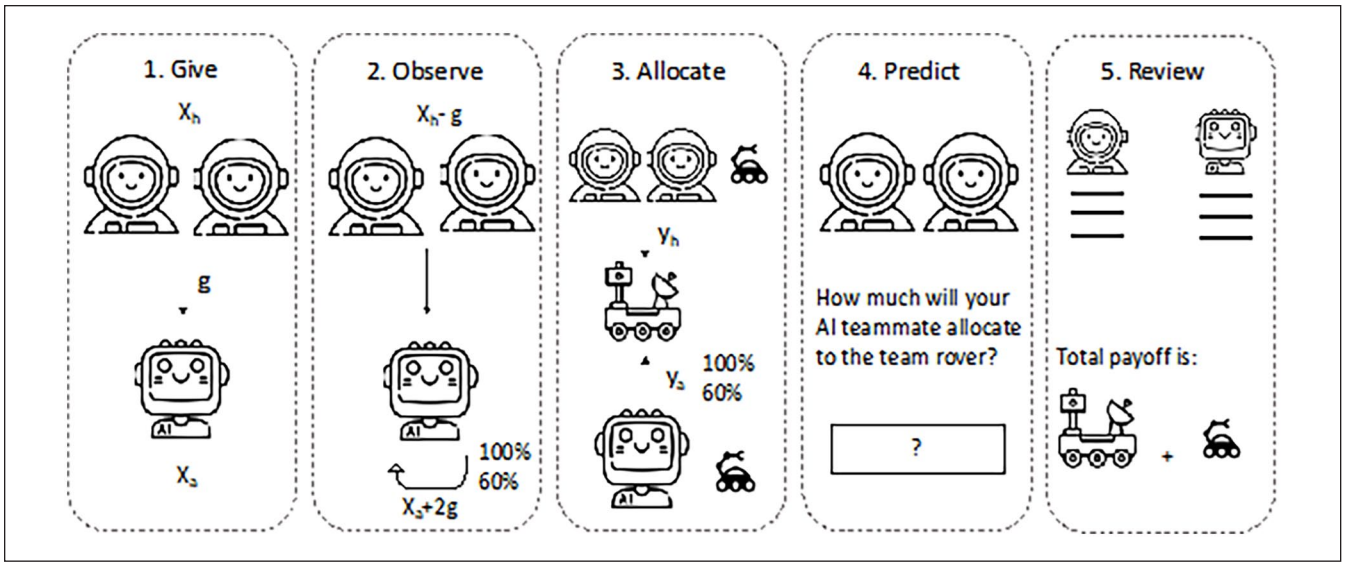

Figure 1. Procedure for space exploration game.

freely discussed their strategies and their perception about the AI teammate's performance. During the beginning and end of each round, the confederate made pre-trained and consistent utterances based on a script that includes key words for each level of trust. Their conversations were recorded using two microphones and a skeleton-based camera. By the end of the first, fifth, and tenth rounds, the participant, without the confederate observing the participant's ratings, rate the AI teammate and their human partner using the MDMT scale to evaluate the participant's trust on the confederate and on the AI teammate. After the game finished, participants will fill out demographic information, propensity to trust scale, and individualism-collectivism scale. Afterwards, the confederate leaves the room, and a semi-structured interview was conducted asking questions pertaining to how they felt about the game, their AI and human teammates, and if their decisions were influenced by the human teammate. In the end, participants were debriefed on the purpose of the study, informed about the confederate, and compensated.

Data Analysis

Data were exported directly from the Firebase platform and analyzed using R via R studio, using packages *lme4* and *emmeans* (Bates et al., 2015; Searle et al., 1980). The manipulation check was conducted using one-way ANOVA with the post hoc Tukey HSD test. To examinate trust contagion, we fitted linear mixed models (LMM) for trust in AI teammate, trust in human teammate and game scores. Using the likelihood ratio test, we used the best fit model with *Confederate Trusting* and *AI Reliability*, as their interactions as fixed effects, with subject ID as random effect for the following analysis. For the significant effects, pairwise comparisons were conducted using *emmeans* with Bonferroni correction. Additionally, we used the same model to measure the total game score.

Results

Manipulation Check

We first conducted the manipulation check on participants' perceived confederate's trust in AI teammate to verify the manipulated confederate's trust toward the AI teammate were properly recognized. The one-way ANOVA found a significant effect in *Confederate Trusting* levels for the manipulation check, $F(2, 123) = 63.91$ $p < .001$, $\eta 2 = 0.51$. A Tukey HSD test verified that people rated their perceived trust in confederate significantly higher when interacting with high-trusting confederate ($M = 6.45$, $SD = 0.47$), comparing to the low-trusting condition ($M = 2.93$, $SD = 1.61$), $t(123) = 11.29$, $p_{adj} < .001$. Similar significant effects were found for high-neutral, $t(123) = 6.18$, $p_{adj} < .001$, and low-neutral condition comparisons, $t(123) = 5.15$, $p_{adj} < .001$. These results suggest that the manipulation check for confederate trusting levels toward the AI teammate was properly recognized across all three conditions.

Trust in AI Teammate

The main effect of *AI Reliability* is statistically significant and negative, $\beta = -1.73$, $t(118) = -4.92$, $p < .001$. Participants dropped their trust in AI significantly when interacting with a low-reliability AI teammate, $p_{adj} = .003$. The main effect of *Confederate Trusting* Condition is statistically significant and negative, $\beta = -.76$, $t(118) = -2.19$, $p = .031$. Participants interacting with the high-trusting confederate showed

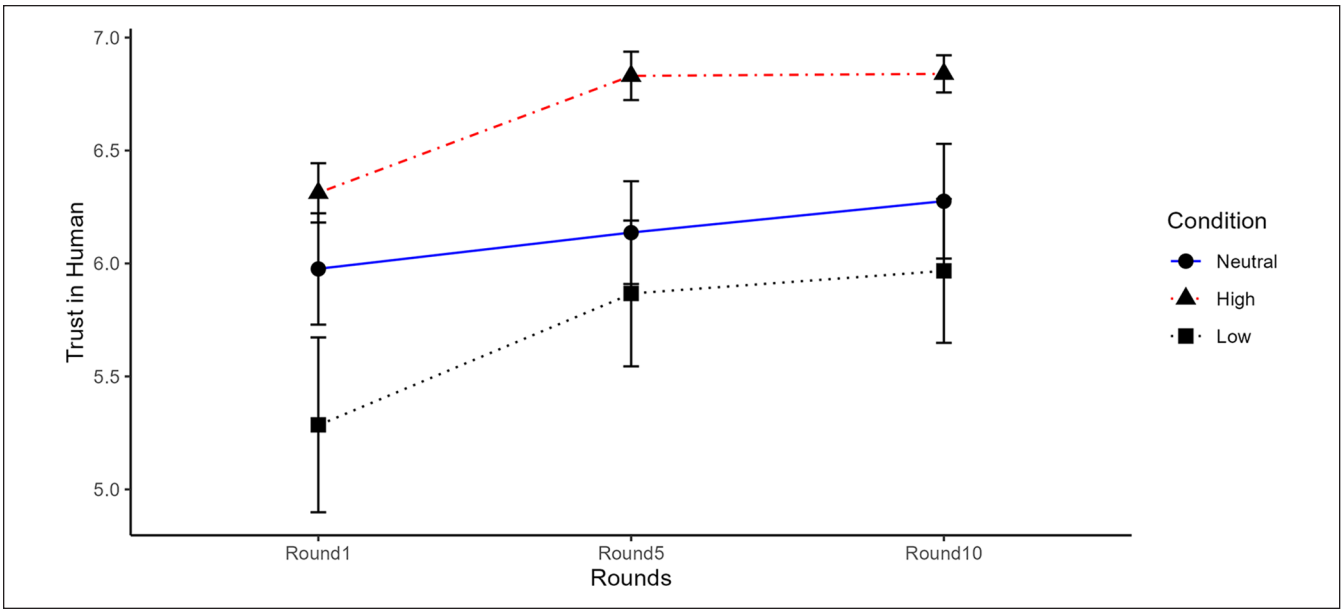

Figure 2. Trust in AI teammate in each round between confederate's trusting conditions.

significantly higher trust in the AI than neutral (p_{adj} = .012) and low conditions (p_{adj} < .001). The interaction effect of *Confederate* Trusting Condition [high] and *AI Reliability* [low] is also statistically significant and positive, β = 1.46, $t(118) = 2.93$, $p = .004$. As shown in Figure 2, when interacting with a high-trusting confederate, participants scored trust in AI significantly higher than neutral (p_{adj} = .004) and low condition (p_{adj} = .002) in round 10. This conveys that there is evidence in the high-trusting confederate enacted positive contagion of the AI trust to the participant, even after the low reliability rounds.

Trust in Human Teammate

The main effect of *AI Reliability* is statistically significant and positive, β = .39, $t(118) = 2.45$, $p = .016$. Participants significantly increased their trust in the confederate during the low-reliability rounds, p_{adj} = .002. The main effect of *Confederate Trusting* is statistically significant and positive, β = .99, $t(118) = 2.95$, $p = .004$. As shown in Figure 3, participants interacting with the high-trusting confederate showed significantly higher trust in the confederate than the low condition, p_{adj} = .022. However, no significant difference between the neutral and low *Confederate Trusting* conditions was found, $p_{adj} > .05$. Additionally, there was no significant difference between the neutral and high *Confederate Trusting* conditions, $p_{adj} > .5$.

Game Performance

For the model fitting for the game performance, we excluded *AI Reliability* in the model because the game score is accumulative and only reported at the end of round 10. The main effect of *Confederate Trusting* condition is statistically significant and negative, β = −125.57, $t(39) = -3.55$, $p = .001$. Participant dropped their total game scores significantly in the low *Confederate Trusting* condition than the neutral (p_{adj} = .002) and high (p_{adj} < .001) conditions, conveying the confederate enacted negative trust contagion to make the participant distrust the AI leading to a low total game score.

Discussion

In this study, we introduced and defined the concept of trust contagion in a human-human-AI team. We explored how the trust levels of one human teammate in AI influences the trust and reliance behaviors of the other human teammate. Our results showed that trust is indeed contagious in human-AI teams. Specifically, when the human teammate indicated a high trust in the AI teammate, participants also trusted the AI teammate more and performed better by relying more on AI teammate in the task. These results align with group emotional contagion effects, which further supports that trust is an affective-laden process. Our research also extends the understanding of trust from dyadic teams to multi-human teams, showing how trust is influenced not only by characteristics of AI teammates but also by social influences from human teammates.

Our results did not support the second hypothesis that distrust would be more contagious than trust. This aligns with Barsade's (2002) study, which showed the contagion of negative mood was not more powerful in inducting emotion than the positive mood. One possible explanation is the nature of the cooperative joint- decision space exploration game in our experiment, where cooperation is critical for success, making positive affect and trust direction more susceptible to participants. Future studies can further investigate the

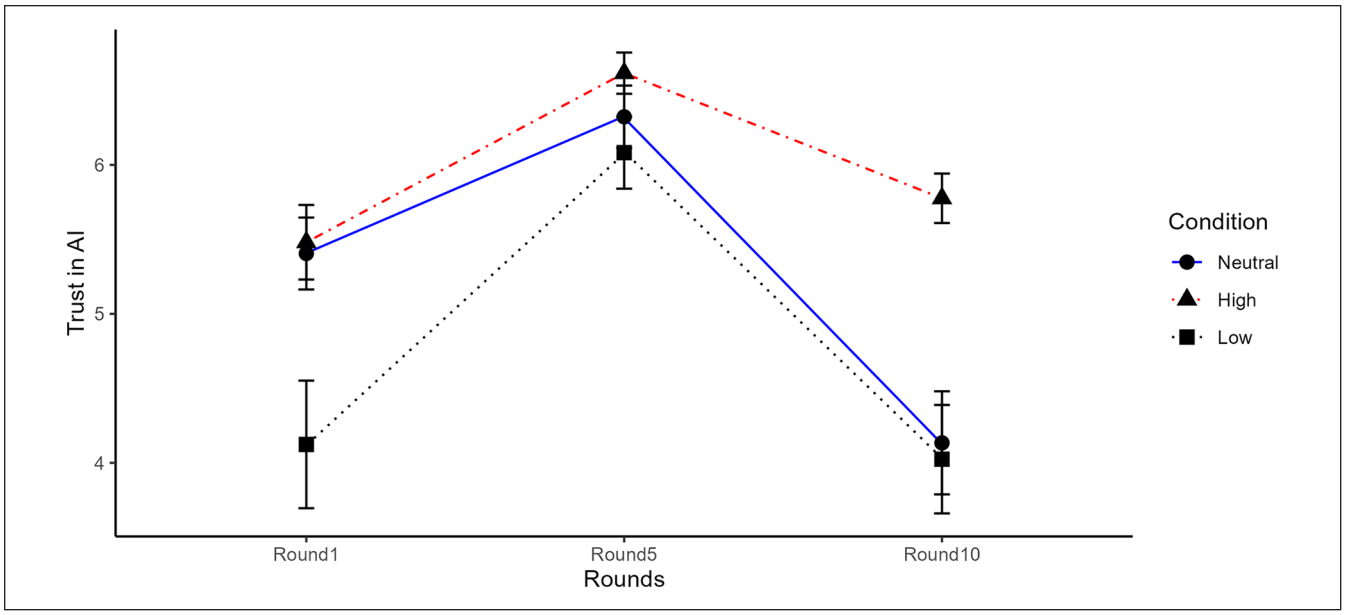

Figure 3. Trust in confederate in each round between confederate's trusting condition.

contagion of both trust and distrust in a neutral setup to further explore their effects.

Since the trust contagion effect does occur in human-AI teams, future studies can further model these social influences using more nuanced physiological and behavioral data to identify specific verbal cues or non-verbal cues (Pantic et al., 2011). In particular, a mapping of emotional valence from their facial expressions, gaze behaviors, and body language can potentially convey social signals of trust contagion. Findings from the social signal processing and modeling can guide design of AI agent's countermeasures to mitigate any inappropriate contagion effects in the team, facilitating the trust calibration process.

Acknowledgments

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

Declaration of Conflicting Interests

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

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

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.

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.

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.

Feese, S., Arnrich, B., Tröster, G., Meyer, B., & Jonas, K. (2012). Quantifying behavioral mimicry by automatic detection of non-verbal cues from body motion. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 520–525.

Guo, Y., Yang, X. J., & Shi, C. (2023). TIP: A trust inference and propagation model in multi-human multi-robot teams. *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, 639–643.

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

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, Augmented and Mixed Reality. Applications and Case Studies* (pp. 476–489). Springer International Publishing.

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

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In *Trust in Human-Robot Interaction* (pp. 3–25).

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.

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.

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.

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