

Compromise in Human-Robot Collaboration for Threat Assessment

Gerald Matthews¹, Ryon Cumings¹, James Casey²,
April Rose Panganiban³, Antonio Chella⁴, Arianna Pipitone⁴,
Jinchao Lin⁵, and Mustapha Mouloua⁵

Abstract

Advancements in Artificial Intelligence (AI) will produce “reasonable disagreements” between human operators and machine partners. A simulation study investigated factors that may influence compromise between human and robot partners when they disagree in situation evaluation. Eighty-seven participants viewed urban scenes and interacted with a robot partner to make a threat assessment. We explored the impacts of multiple factors on threat ratings and trust, including how the robot communicated with the person, and whether or not the robot compromised following dialogue. Results showed that participants were open to compromise with the robot, especially when the robot detected threat in a seemingly safe scene. Unexpectedly, dialogue with the robot and hearing robot inner speech reduced compromise and trust, relative to control conditions providing transparency or signaling benevolence. Dialogue may change the human’s perception of the robot’s role in the team, indicating a design challenge for design of future systems.

Keywords

decision-making, human-robot interaction, threat assessment, trust, transparency

Introduction

Applications for collaboration between humans and intelligent robots and virtual agents are rapidly proliferating. The resolution of conflicts is critical for successful teaming; conflicts can arise from team members having different perspectives on both task requirements and the person’s role in the team (de Wit et al., 2013). In human-autonomy teaming, conflict between human and AI-powered collaborators may escalate as machines take over functions previously reserved for humans, and as human-agent teams are increasingly deployed for complex, open-ended missions in dynamic, unpredictable environments. Humans and machines may arrive at differing reasoned evaluations due to their differing analytic competencies and situation understanding. Resolving these “reasonable disagreements” burdens the human with understanding machine intelligence and its limitations in order to calibrate trust and optimize decision-making. The present study used a simulation methodology to explore factors that influence conflict resolution during human-AI collaboration to evaluate threat.

The study built on our previous simulation-based research on trust in collaborative threat assessment. Novel factors such as AI explainability and the human’s mental model of the robot influence trust over and above conventional influences such as robot performance (Lin et al.,

2022; Matthews et al., 2024). Figure 1 shows the conceptual model developed from this work. As in human-human teams, constructive dialogue between human and robot partners is critical for optimizing trust and arriving at evidence-based compromise when conflicts arise. The increasingly humanlike capabilities of robots (Wynne & Lyons, 2018) can support communication and teamwork behaviors necessary to support dialogue.

The design challenge is to identify robot attributes that can be manipulated to promote constructive communication. Existing research has focused on transparency: additional information that supports human understanding of robot situation assessment, reasoning and future projection (Chen et al., 2018). Transparency can enhance trust calibration, performance, and situation awareness. As well as

¹George Mason University, Fairfax, VA, USA

²Virginia Serious Games Institute, George Mason University, Fairfax, VA, USA

³Air Force Research Laboratory, OH, USA

⁴Università degli Studi di Palermo, Italy

⁵University of Central Florida, Orlando, FL, USA

Corresponding Author:

Gerald Matthews, Department of Psychology, George Mason University, Fairfax, VA 22030-4444, USA.

Email: gmatthe@gmu.edu

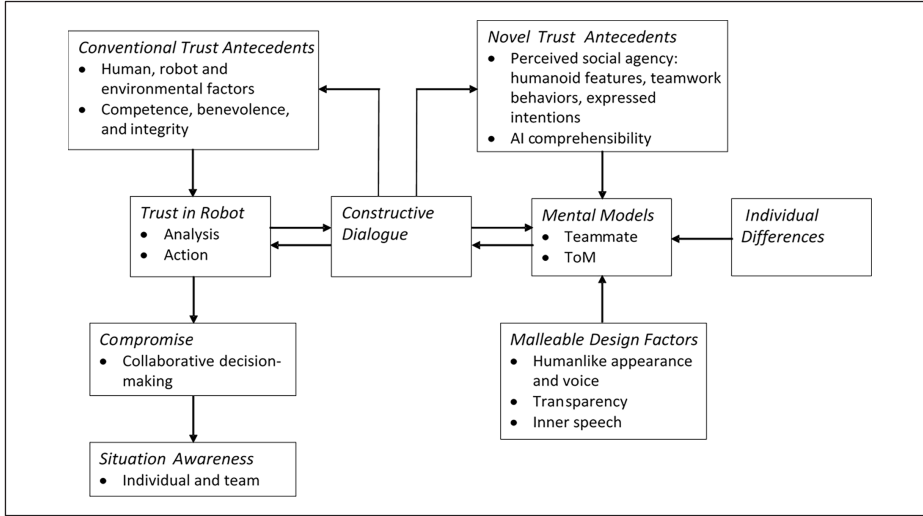


Figure 1. Conceptual framework for compromise in human-robot interaction.

enhancing trust calibration and performance, transparency may also facilitate human-robot communication, although empirical findings are mixed (Wright et al., 2022). Beyond transparency, human willingness to engage with a robot in a conflict situation may depend on the human’s mental model of the robot as a teammate (Matthews et al., 2024). Factors promoting a humanlike mental model include anthropomorphic qualities (Wynne & Lyons, 2018) and attribution of a Theory of Mind (ToM) to the robot (Vinzani et al., 2019). Pipitone et al. (2023) highlighted the role of inner speech as a window in robot mentalizing. They conducted an experiment in which inner speech during human-robot interaction was generated by a cognitive architecture. Inner speech was found to enhance perceptions of the robot and trust, implying that it might also contribute to conflict resolution, communicating to the human that the robot has reasonable grounds for disagreement.

The present study required the participant to partner with an intelligent robot to assess the threat level of 24 urban scenes containing multiple animated human figures and other objects. The participant viewed an initial text briefing on possible threats. The participant made an initial threat rating, received a report from the robot partner based on its sensor analysis, and updated their threat rating if they wished. Half the scenes were designed to look safe when the participant first rated them, based on their briefing and visible elements. Half were designed to appear threatening though inclusion of salient cues to threat. We designed three types of

robot interaction scenario. In 8 agreement scenarios, the robot provided information consistent with the initial scenario design (safe or threatening). Disagreement scenarios were designed so that the robot had a rational basis for assessing threat level contrary to the participant’s initial impression. The robot might have additional sensor or briefing information, it might misunderstand the context for the scene, or its sensors might be malfunctioning. There were eight disagreement/ compromise scenarios in which the robot would change its threat rating toward the participant’s anticipated response: for example, for a “safe” scene, we expected the participant to initially rate threat as low. The robot was scripted to rate threat as high, but would compromise by lowering its threat rating following dialogue. There were eight disagreement/no compromise scenarios in which the robot would not change its threat rating.

We also manipulated the style of communication with the robot. In two control conditions, the participant either received a general statement of benevolence or a one-time transparency message. In experimental conditions, the participant could choose to dialogue with the robot to question its analysis. One condition allowed dialogue via scripted text messages; the other condition also introduced robot inner speech to highlight its mental processes (Pipitone et al., 2023). The cognitive architecture used in previous studies (e.g., Pipitone & Chella, 2021), appropriately calibrated and modified for this investigation, provided the basis for generating inner speech.



Figure 2. Screenshot showing dialogue between human and robot, threat ratings, and sensor status.

The principal outcome variables were (a) change in threat rating following interaction with the robot and (b) trust in the robot. We expected that disagreement scenarios would induce greater changes in threat ratings than agreement scenarios (human compromises) and loss of trust. We expected that if the robot compromised, the human would be more likely to reciprocate by changing their ratings toward those of the robot's and would show higher trust, compared with disagreement/no compromise scenarios. We also hypothesized that the two dialogue conditions would elevate human compromise (larger changes in threat rating) and trust in the robot, and that these effects would be stronger with robot inner speech. Participants could ask the robot up to six questions in each scenario. We expected that inner speech would increase the number of questions asked.

Method

Participants

Participants were 87 psychology students (49 male, 38 female) at a mid-Atlantic state research university. Age range was 18 to 50. They received course credit for participation.

Simulation of Human-robot Collaboration

The simulation was modified from the one used by Lin et al. (2022), programmed in Unreal Engine. The participant was instructed to collaborate with a robot partner to evaluate possible threats to a convoy at 24 separate scenes. Figure 2 illustrates the interface for task performance. The left window presented the initial briefing and supported dialogue between participant and robot. The upper right window allowed the participant to enter and modify threat ratings. It also presented the robot's threat ratings. The lower right window indicated which sensors the robot had mounted and their status, including faults. Sensors varied from scene to scene; examples were chemical and metal detectors, speech content analyzers, and off-body physiological sensors to detect anger

and aggression. The participant used a mouse to select threat ratings and pre-scripted dialogue options.

Threat level and scenario type were manipulated as previously described. There were eight agreement scenarios, eight disagreement scenarios in which the robot would compromise following dialogue, and eight no-disagreement scenarios. For each scenario type, there were four scenes designed to appear safe when first viewed, and four designed to appear threatening.

Procedure

Following consent, participants completed a demographics survey, and questionnaires on trust in robots. They were assigned at random to one of four communication conditions (benevolence control, transparency control, dialogue with no inner speech, dialogue with inner speech) and completed training and practice for that condition. They then performed the threat evaluation task for the 24 scenarios, presented in a single pseudo-random order. Questionnaires were administered post-task to assess situational trust and stress (questionnaire data are not reported here). The sequence of events in the dialogue conditions was as follows:

1. Participant views mission brief and scene, makes initial threat rating on a 1 to 5 scale.
2. Robot provides threat rating.
3. Robot provides report with threat rating and explanation.
4. Participant can choose to modify their threat rating.
5. Participant can choose to question robot. There were two initial questions available, plus two follow-up questions after each initial one.
6. Robot provides final report; may (or may not) revise its threat ratings.
7. Participant can choose to modify their threat rating.
8. Post-scene ratings of confidence in robot analysis and robot recommendation (1–5 scales). Emotions were also rated.

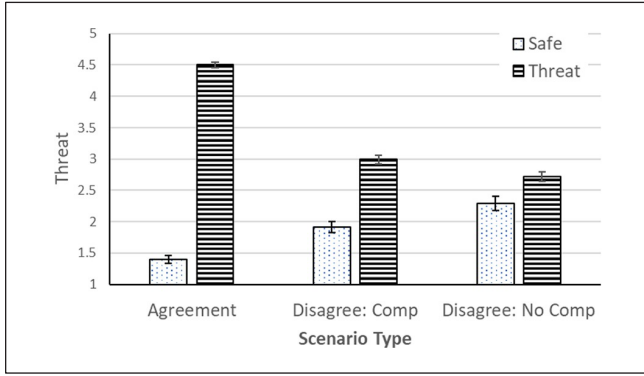


Figure 3. Mean threat rating as a function of scenario type and scenario threat (Comp = Compromise).

The participant heard robot inner speech at step #5 in the inner speech condition. In the control conditions, there was no dialogue and steps #5 to 7 were omitted. In these conditions, at step #3 the robot either provided a short text justification for its evaluation (transparency) or made a generic statement supportive of the human (benevolence).

Results

Threat Ratings

The simulation logged ratings prior to and following interaction with the robot (initial and final threat ratings). We confirmed that initial mean threat ratings were higher in threat scenarios (mean = 3.32; $SD = .47$) than in safe scenarios (mean = 1.51; $SD = .29$). To examine effects of scenario characteristics on final threat ratings, we ran a repeated-measures 3×2 (scenario type $\times$ scenario threat) ANOVA. There were significant main effects of scenario type, $F(2, 172) = 62.11, p < .001, \eta_p^2 = .419$, and scenario threat, $F(1, 86) = 674.81, p < .001, \eta_p^2 = .887$, as well as a type $\times$ threat interaction, $F(2, 172) = 382.21, p < .001, \eta_p^2 = .792$. These effects are illustrated in Figure 3. In agreement scenarios, participants saw safe scenarios as very safe and threat scenarios as very threatening, as expected. In disagreement scenarios, the contrast between safe and threat scenarios was much reduced, especially if the robot did not compromise. Impacts of disagreement were stronger for threat than for safe scenarios. Participants appeared to compromise with the robot, even (or especially) if the robot did not itself compromise.

Analysis of robot communication focused on changes in ratings following interaction with the robot. We analyzed Δ Threat, that is, final threat rating—initial threat rating, using a mixed-model $3 \times 2 \times 4$ (scenario type $\times$ scenario threat $\times$ communication) ANOVA, with repeated measures on the first two factors. Consistent with the preceding ANOVA, there were significant effects of scenario type, $F(2, 166) = 10.45, p < .01, \eta_p^2 = .112$, scenario threat, $F(1, 83) = 103.38, p < .001, \eta_p^2 = .555$, and the

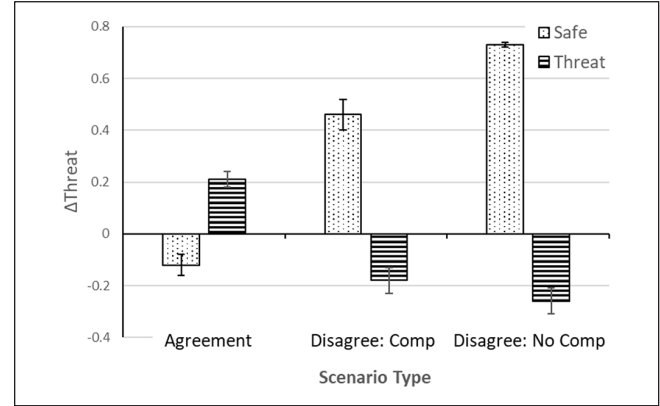


Figure 4. Mean change in threat rating as a function of scenario type and scenario threat (Comp = Compromise).

interaction, $F(2, 166) = 96.43, p < .001, \eta_p^2 = .537$, as shown in Figure 4. In agreement conditions, interaction with the robot reinforced the person's initial evaluation: safe scenarios were seen as slightly less threatening and threat scenarios as slightly more threatening. These trends were reversed when the robot disagreed. Participants compromised with the robot by elevating their threat ratings for safe scenarios and depressing their ratings of threat scenarios. These trends were stronger for safe scenarios and where the robot did not compromise.

The main effect of communication was nonsignificant, but there were significant scenario threat $\times$ communication, $F(3, 83) = 10.65, p < .001, \eta_p^2 = .279$, and scenario type $\times$ scenario threat $\times$ communication, $F(6, 166) = 4.52, p < .001, \eta_p^2 = .140$, interactions. Follow-up analyses conducted for each threat level independently showed significant main effects of communication for both safe scenarios, $F(3, 83) = 4.70, p < .001, \eta_p^2 = .145$, and threat scenarios, $F(3, 83) = 6.97, p < .001, \eta_p^2 = .201$. The scenario type $\times$ communication interaction was close to significance for safe scenarios, $F(6, 166) = 2.06, p = .060, \eta_p^2 = .069$, and significant for threat scenarios, $F(6, 166) = 3.84, p = .01, \eta_p^2 = .122$. Figure 5 illustrates the interactions. In the two control conditions, participants compromised substantially. They moved their ratings toward those of the robot, elevating threat rating where the robot saw threat, and lowering their rating where the robot identified a safe situation. By contrast, in dialogue conditions, participants stayed closer to their initial ratings, especially in the compromise condition.

Trust

Trust was assessed as the mean of the two ratings for confidence in the robot's analysis and acceptance of its recommendations. This trust score was analyzed using a mixed-model $3 \times 2 \times 4$ (scenario type $\times$ scenario threat $\times$ communication) ANOVA. There were significant effects of scenario type, $F(2, 166) = 286.91, p < .001, \eta_p^2 = .776$, scenario threat,

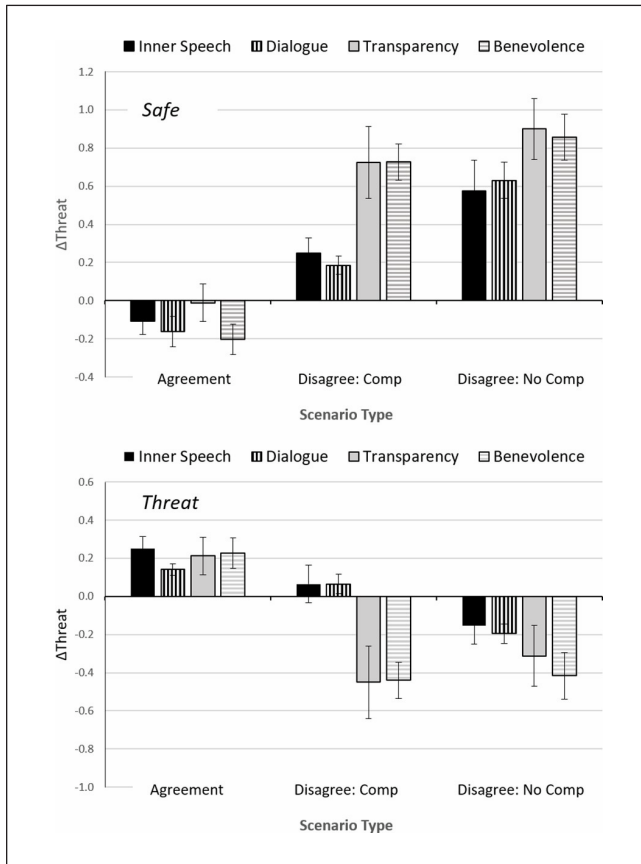


Figure 5. Δ Threat as a function of scenario and communication condition for safe (top) and threat scenarios (bottom). (Comp = Compromise).

$F(1, 83) = 69.07, p < .001, \eta^2_p = .454$, and the interaction, $F(2, 166) = 8.72, p < .001, \eta^2_p = .095$.

Trust was higher for agreement than for disagreement scenarios (see Figure 6). In the latter case, participants trusted the robot more if it identified a seemingly threatening scenario as safe than if it rated a safe scenario as threatening. Whether or not the robot compromised had little effect on trust.

There was also a main effect of communication, $F(3, 83) = 6.23, p < .001, \eta^2_p = .184$, moderated by a scenario threat $\times$ communication interaction, $F(3, 83) = 2.97, p < .05, \eta^2_p = .097$. Post-hoc Sidak tests showed that the inner speech condition produced lower trust than the two control conditions ($p < .05$). The modest interaction was associated with a stronger effect of communication in scenario threat conditions (η^2_p s: .178 vs. .145), although the effect was independently significant ($p < .01$) at both levels of threat.

Questions Asked

Data on mean frequency of questions were analyzed for the two dialogue conditions, that is, with and without inner speech ($N = 46$). The analysis showed a main effect of

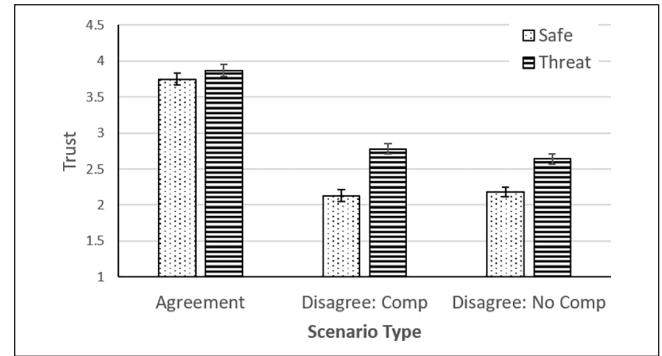


Figure 6. Mean trust rating as a function of scenario type and scenario threat (Comp = Compromise).

scenario type, $F(2, 88) = 8.27, p < .001, \eta^2_p = .324$ and a scenario type $\times$ scenario threat interaction, $F(2, 44) = 6.99, p < .01, \eta^2_p = .137$. There were no significant or main effects of communication. Means (and SDs) were 4.18 (1.44) for agreement scenarios, 4.91 (1.35) for disagreement/compromise and 4.70 (1.36) for disagreement/no compromise. Participants asked more questions in disagreement conditions, as expected. The scenario effect tended to be stronger in threat conditions.

Discussion

Human-robot collaboration in challenging environments may increasingly generate conflict when partners arrive at different situation assessments. Findings from this study showed that human reactions to a robot with contrary views on the level of threat posed by a scene depended on multiple factors. These included initial perceptions of threat or safety, whether the robot compromised, and the nature of communication between human and robot. Several effects on threat rating and trust were consistent with expectation. In agreement scenarios, in which the robot's report was consistent with the cues afforded by the visible scene and the participant's brief, people straightforwardly assigned high or low threat ratings consistent with the scenario design. Participants were also sensitive to the robot reports in disagreement scenarios, showing a broad tendency toward compromising with the robot by lowering their threat ratings when the robot reported that the scene was safe and elevating ratings when the robot perceived threat. Similarly, Lin et al. (2022) found that people seek to rationally integrate information from multiple cues to evaluate threat. In agreement scenes, people also expressed moderately high trust in the robot, contrasting with lower trust ratings when the robot disagreed with them.

Other features of the data were not anticipated. We predicted that participants would react positively to robot compromise in cases of disagreement, compromising themselves and trusting the robot more than in no-compromise scenarios. In fact, we saw the opposite trend; people shifted their

ratings more toward the robot's evaluation when the robot did not compromise. In the disagreement scenarios, compromise by the robot also failed to increase trust. We also observed asymmetry in reactions to scenes designed to appear safe versus threatening initially. In disagreement scenarios, participants showed larger-magnitude compromise for threat scenarios and assigned higher trust to the robot, relative to safe conditions. These findings may partly reflect a rational response to the robot's capacity to detect cues to threat imperceptible to the human, such as chemical traces. By contrast, in threat scenarios, people may be reluctant to reverse their initial impression of danger following the robot report.

Effects of robot communication style were also contrary to the initial hypothesis. We expected that dialogue, and especially dialogue with inner speech, would promote compromise with the robot. In fact, we found the opposite. In cases of disagreement, participants in the dialogue conditions changed their ratings only minimally following interaction with the robot, whereas those in control conditions shifted their ratings substantially. Dialogue, especially with inner speech, also reduced trust in the robot. These findings contrast with those of Pipitone et al. (2023) who found that inner speech enhanced engagement with the robot. There are several possible reasons for the discrepancy. First, Pipitone et al. (2023) utilized a physical, somewhat humanlike robot. Inner speech effects may be less effective in conveying humanlike mentation in a virtual agent. Second, Pipitone et al.'s task did not directly generate conflict. Third, in the present context, dialogue with the robot may reveal its limitations, especially at the beginning of interactions, eroding trust. Fourth, we assumed that dialogue and inner speech might illuminate the robot's taskwork. However, it may also have impacted the participant's sense of the robot's role in the team (de Wit et al., 2013), causing it to be seen as a rival or collaborator rather than a cooperating teammate (see also Pipitone et al., 2023).

From a practical standpoint, results are encouraging in showing that, at least in novices, humans are broadly willing to compromise with a robot even when it disagrees with them and does not itself compromise. Acceptance of robot analyses may be necessary in cases where the AI output is not readily explainable, including operational situations in which there is high time pressure to respond (e.g., Lyons & Guznov, 2019). Interestingly, people are willing to compromise even when trust in the robot is relatively low. Further work might address design features that optimize compromise and trust, which would require objective standards lacking in this study. Design should address separately mitigation of "missed targets" (failure to recognize threat) and "false positives" (attributing threat to safe scenarios). In security or military settings, these errors may have different consequences and AI should guide the operator toward evaluating safety and threat accordingly.

A final practical issue is the design of bilateral robot communication (Wright et al., 2022). Participants typically asked multiple questions of the robot even in agreement scenarios suggesting that communication was engaging for them. However, dialogue did not facilitate compromise or increase trust. Future work should address how design of communication should support optimization of compromise and trust against performance standards. It should also address how dialogue and inner speech can support both the taskwork and team role elements of conflict resolution (de Wit et al., 2013). In cases where the person must rely on robot analysis, even if it is incomprehensible, the role of dialogue may be to assure the person that the robot remains a good team player, even though its recommendations differ from those of the human. In cases where the human and the robot must pool their differing knowledge expertise, dialogue should actively facilitate compromise with a focus on resolving different perspectives on task accomplishment.

Limitations of the study include use of a simulated environment and novice participants. Professional security agents with expertise in threat detection may have been less willing to compromise with the robot. The study was designed to investigate relative levels of threat rating and trust in the different scenarios. However, with no objective standard for threat, responses could not be evaluated for threat detection accuracy or trust optimization. A further limitation is that scenarios were hand-crafted so that results might be influenced by idiosyncratic features of the design. In addition, the robot disagreed with the human in the majority of scenarios; results may have differed if the robot typically agreed with the human.

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) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: We gratefully acknowledge support from the Air Force Office of Scientific Research (AFOSR) Trust and Influence program (award FA9550-22-1-0035).

References

- Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, 19(3), 259–282.
- de Wit, F. R., Jehn, K. A., & Scheepers, D. (2013). Task conflict, information processing, and decision-making: The damaging effect of relationship conflict. *Organizational Behavior and Human Decision Processes*, 122(2), 177–189.
- Lin, J., Panganiban, A. R., Matthews, G., Gibbins, K., Ankeney, E., See, C., Bailey, R., & Long, M. (2022). Trust in the

-
- danger zone: Individual differences in confidence in robot threat assessments. *Frontiers in Psychology*, 1426, 601523.
- Lyons, J. B., & Guznov, S. Y. (2019). Individual differences in human-machine trust: A multi-study look at the perfect automation schema. *Theoretical Issues in Ergonomics Science*, 20(4), 440-458.
- Matthews, G., Hancock, P., Szalma, J., Lin, J., & Panganiban, A.R. (2024). Individual differences in teaming with artificial intelligence, robots, and virtual agents in the workplace. In A.A. Oguz, A. Tuncdogan, H. Volberda, & K. de Ruyter (Eds.), *The Oxford handbook of individual differences in organizational contexts*. Oxford: Oxford University Press.
- Pipitone, A., & Chella, A. (2021). What robots want? Hearing the inner voice of a robot. *Iscience*, 24(4), 102371.
- Pipitone, A., Geraci, A., D'Amico, A., Seidita, V., & Chella, A. (2023). Robot's inner speech effects on human trust and anthropomorphism. *International Journal of Social Robotics*, 16, 1333-1345.
- Vinanzi, S., Patacchiola, M., Chella, A., & Cangelosi, A. (2019). Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B*, 374(1771), 20180032.
- Wright, J. L., Lakhmani, S. G., & Chen, J. Y. (2022). Bidirectional communications in human-agent teaming: The effects of communication style and feedback. *International Journal of Human-Computer Interaction*, 38(18-20), 1972-1985.
- Wynne, K. T., & Lyons, J. B. (2018). An integrative model of autonomous agent teammate-likeness. *Theoretical Issues in Ergonomics Science*, 19(3), 353-374.