Trust in Autonomous Human–Robot Interaction

An In-Person Pilot Study

M.C. Lau
Laurentian University
mclau@laurentian.ca

Shauna Heron
Laurentian University
sheron@laurentian.ca

2025-12-14

Abstract This study implements a multi-stage collaborative task system where participants collaborate with the Misty II social robot to solve a who-dunnit type task. The system utilizes an autonomous, mixed-initiative dialogue architecture with affect-responsive capabilities.

Human–Robot Collaboration (HRC) has emerged as a critical area in the engineering and social sciences domain. The current paper ventures into this growing domain with a focus on an environment where collaboration pivots on problem-solving through shared knowledge and dialogue. Specifically, we explore collaboration performance and trust perception after interaction with two versions of a social robot capable of carrying out autonomous actions and decision-making under the guidance of a computer program. One version designed to be responsive and proactive to participants affect and queries and the other designed to offer help only when asked.

Trust is a central construct in human–robot interaction (HRI), shaping how people collaborate with, rely on, and accept robotic systems across social, assistive, and task-oriented domains [1]. In any kind of collaborative setting, including HRI, trust has been identified as a significant factor that can work to support or hinder cooperation, particularly in contexts characterized by incomplete or uncertain information [2]. Trust influences not only subjective evaluations of robots but also objective outcomes like task performance, compliance, and engagement [3]. As a result, a growing body of work has focused on evaluating trust following human–robot interactions, including the development of several standardized instruments designed to capture users' perceptions of robot reliability, predictability, and intent in various industrial, medical and social settings [2], [3], [4], [5].

Despite this growing literature, much of what is currently known about trust in HRI has been derived from interactions conducted under simulated conditions. In many studies, robot behavior is scripted, simulated by a computer program, or mediated through human control of a robot using Wizard-of-Oz (WoZ) paradigms [6], [7]. While such approaches are valuable for early-

stage design and hypothesis generation, these approaches critically alter interaction dynamics by masking real-world sensing failures, response latency, and behavioral inconsistencies that are characteristic of autonomous robotic systems. This gap is especially notable given that autonomy-related challenges—such as speech recognition errors, model hallucinations, delayed responses, and misinterpretations of user intent—are likely to play a critical role in shaping trust during real deployments. From an HRI perspective, understanding trust in the presence of real-world imperfections may be more informative than evaluations conducted under idealized assumptions. Nevertheless, few studies have directly examined trust outcomes following fully autonomous, in-person human–robot interaction.

To address this gap, the current study leveraged a between-subjects design to evaluate trust following an in-person interaction with a robot operating autonomously within predefined behavioral constraints. Participants collaborated with the robot in solving an immersive puzzle game where the robot served as a diegetic "game guide" and collaborative partner. In the game, the participant solved the mystery by interacting with the game guide to obtain hints, moral support and advice on how to solve puzzles; the robot managing speech-based interaction, task progression, and affect-responsive behavior, all without human intervention.

In the control condition the robot followed a neutral interaction policy, while in the experimental condition the robot was prompted to adapt its behavior based on detected user affect, dialogue and demands of the task itself. Importantly, both conditions utilized the same robot subject to the same sensory and interaction limitations inherent to autonomous operation, including speech recognition variability and response timing constraints. The only difference was the interaction policy between conditions.

To achieve this we developed an autonomous spoken-language interaction system integrated with a speech recognition (ASR) pipeline, including affect detection with the Misty-II robot platform to allow the robot to engage in natural conversations with users. The system is capable of recognizing speech, managing dialogue, remembering what was said previously, and generating spoken responses and facial expressions and head and arm movement of the robot during dialogue.

By examining post-interaction trust using established trust measures alongside behavioral and task-level outcomes, this study aims to contribute empirical evidence on how trust might be shaped in fully autonomous HRI scenarios. Rather than seeking to demonstrate optimal performance under ideal conditions, the focus is on understanding trust as it is impacted during realistic human–robot interaction, where uncertainty, interactional breakdowns, and adaptive behavior are unavoidable. As such, this work provides insight into the practical implications of affect-responsive autonomy for trust in human–robot collaboration.

Task Design

Participants collaborated with the robot in solving an immersive puzzle game where the robot served as a diegetic "game guide" and collaborative partner. In the game, the participant solves the mystery by interacting with the robot guide for hints, emotional support and advice on how to solve the puzzles. The game was composed of two sequential tasks designed to elicit interaction with the robot under differing knowledge and dependency conditions [7] . The robot

autonomously monitored task progression through the interface and adapted its dialogue accordingly without real-time human intervention. The tasks were structured as follows:

Task 1: Robot-dependent collaborative reasoning
The first task required participants to identify a suspect from a 6x4 grid of 'suspects' by asking a series of yes/no questions about their features. A grid of potential suspects was displayed on the interface, and participants formulated questions verbally to narrow down the correct individual (e.g., 'was the suspect wearing a hat?'). In this task, the robot possessed the ground-truth information necessary to determine whether each question was true or false, making successful task completion dependent on interaction with the robot.

This task was designed to establish an initial forced collaborative dynamic in which the robot served as an essential informational partner. Participants were required to engage verbally with the robot and coordinate question strategies to reach a solution within the allotted time (5 minutes). The structured nature of the task ensured that the robot's role was clear and that collaboration was unavoidable.

Task 2: Open-ended problem-solving with advisory robot support
The second task involved a more complex problem-solving scenario in which participants had access to multiple technical logs presented via a simulated terminal interface to determine the location of the missing 'Atlas' robot. Unlike the first task, the robot did not possess ground-truth knowledge about the whereabouts of the robot. The robot's assistance in this task was limited to general problem-solving support derived from language model's prior training, such as explaining how to interpret log information, suggesting reasoning strategies, or helping participants reflect on inconsistencies across logs. The robot was explicitly constrained such that it was informed only that participants could view several logs, without access to the content of those logs or the correct answers to the task-related questions. The robot could ask the participant questions about what they found in the logs and the human could do the same.

Importantly, participants could complete these tasks independently or choose to solicit assistance from the robot. As a result, the robot functioned as a collaborative reasoning partner rather than an authoritative source. Participants retained full control over decision-making and were free to accept, reject, or ignore the robot's suggestions. This design allowed collaboration to emerge voluntarily, rather than being enforced by task structure [7].

When T.-H. Lin, S. Ng, and S. Sebo [7] et al., utilized a similar task they found that participants who engaged with a robot compared to a human guide had more fun, felt less judged and more connected with the robot while solving tasks compared to a human–though respondents mentioned that it would be helpful if the robot was more proactive in the help it provided. Though they utilized a WOz system, in our case, both tasks were completed autonomously in the presence of a shared task interface that displayed instructions, task materials, and participant inputs.

Once all answers were submitted, the correct answers were shown to participants, letting them know how they did. At the wrap-up stage the robot and the participant had the chance to debrief on whether they were right or not, and then the task came to an end with the robot thanking them and prompting them to report to the principal investigator.

Task difficulty and collaborative intent

The second task was intentionally designed to be sufficiently challenging that completing it within the allotted time was difficult without assistance. This ensured that interaction with the robot represented a meaningful opportunity for collaboration rather than a trivial or purely optional exchange. By contrasting a robot-dependent task with an open-ended advisory task, the study examined trust formation across interaction contexts that varied in both informational asymmetry and reliance on the robot.

Across both tasks, the interface served as a shared workspace facilitating coordination between the participant and the autonomous robot, rather than as a mechanism for remotely controlling robot behavior. At no point during either task did a human operator intervene to guide the robot's actions or manage task flow.

System overview and experimental setup

Participants interacted with the Misty-II robot in a shared physical workspace that included both the robot and a computer-based task interface. The interface was visible to participants and used to present brief task instructions, collect responses, and advance between task stages. Importantly, the robot autonomously monitored task progression and participant input through the interface, allowing it to adapt its dialogue and responses without human intervention.The interface served as a communication channel between the participant and the autonomous system rather than as a mechanism for remotely controlling robot behavior (See Figure 1) .

Experimental setup and interaction environment

Participants (n = 29) completed a pre-interaction questionnaire on Qualtrics where consent, demographics information, and a baseline measure of Negative Attitudes Towards Robot Scale and Need for Cognition (thinking style) were administered. Because of potential variability around timing of the pre-interaction tests and the in-person sessions, we elected not to use a formal pre-post test. Instead we took a general measure of attitudes towards robots as well as general thinking style to establish a baseline for later group comparison.

At the in-person session, once the pre-interaction survey was complete, participants were seated in front of Misty and instructed to start the session by clicking the Start button on the dash. They were also instructed on basic communication tips with the robot: i.e., to wait until the blue light on the side of the robot's head was on before speaking. Finally, once the participant was ready to start, the researcher left the room and closed the door, leaving the robot and participant to complete the tasks together. Once complete the participant would exit the room and then complete a post-interaction survey containing the Trust Perception-HRI scale and the Trust in Industrial Human-robot Collaboration scale followed by a written debriefing and verbal debriefing with the primary researcher. Participants were informed they could leave the room and stop the session at any time, no questions asked. Once complete, participants were presented with a $15.00 gift card as compensation for their time. All participants completed the tasks and remained for their full session which took an average of 30 minutes to complete.

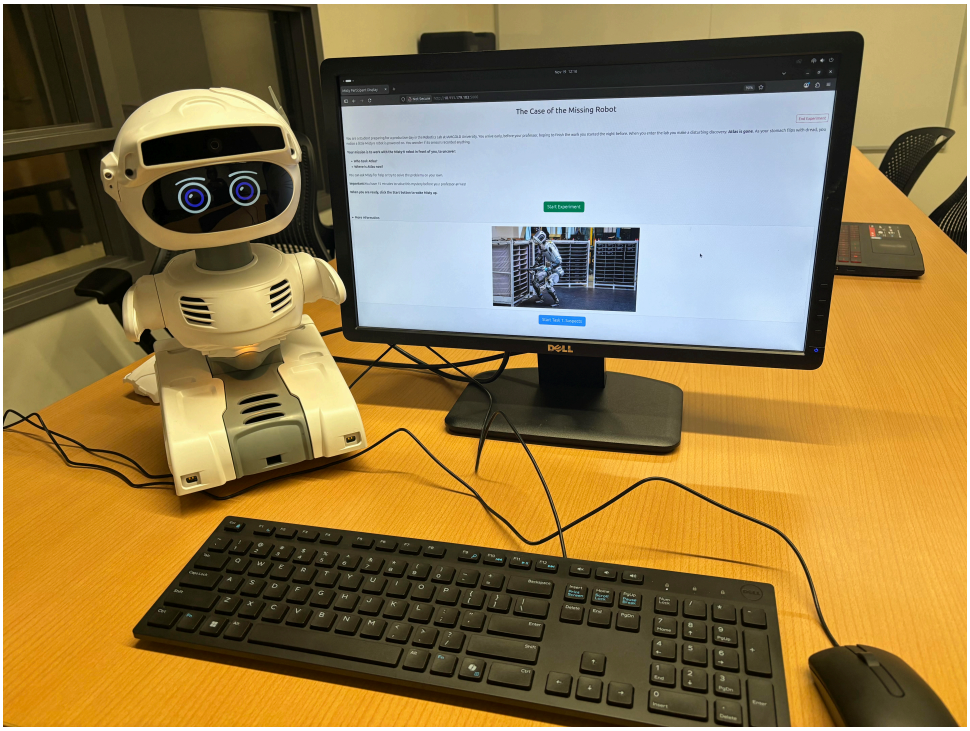

Figure 1: Experimental setup showing the autonomous robot and participant-facing task interface used during in-person sessions. Participants entered task responses and navigated between task stages using the interface, while the robot autonomously tracked task state and adapted its interaction based on participant input. No real-time human intervention occurred during the interaction.

Robotic system and autonomy pipeline

The task interface was adapted from prior work with a similar robotic platform in which a graphical interface was used to support Wizard-of-Oz control T.-H. Lin, S. Ng, and S. Sebo [7] . In the present study, this idea was reimagined as a shared workspace for human–robot collaboration. Rather than serving as a control surface, an interface was developed to function as a shared task environment through which both the participant and the robot maintained awareness of task state and progress. Participant inputs were visible to the robot, allowing it to track task transitions and respond contextually, while all behavioral decisions were generated autonomously by the robot.

The human participant could choose to work independently or solicit assistance from the robot, which provided guidance (both conditions), clarification (both conditions), and proactive, affective support (responsive condition only) but no definitive answers. This task was intentionally designed to be sufficiently challenging that completing it within the allotted time was difficult without assistance, thereby creating a meaningful opportunity for collaboration rather than a trivial interaction. The robot's assistance was framed as collaborative support rather than authoritative guidance, and participants were not led to believe that the robot possessed complete or privileged knowledge during the second task. As a result, the robot's role in the second task was that of a collaborative reasoning partner rather than an authoritative source.

Interaction conditions

Describe the responsive versus control conditions here briefly. Maybe give explanation of how that was handled with langchain?

Results

Participant characteristics and baseline measures

Participants in the control and responsive conditions were comparable with respect to demographic characteristics, academic background, prior experience with robots, and baseline attitudes toward robots. Importantly, Negative Attitudes Towards Robots (NARS) and Need for Cognition scores were similar across groups, indicating that post-interaction differences are unlikely to reflect pre-existing attitudes (see **?@tbl-pre**).

References

Bibliography

[1] E. Loizaga, L. Bastida, S. Sillaurren, A. Moya, and N. Toledo, "Modelling and Measuring Trust in Human–Robot Collaboration," *Applied Sciences*, vol. 14, no. 5, p. 1919, Jan. 2024, doi: 10.3390/app14051919.

[2] K. E. Schaefer, "Measuring Trust in Human Robot Interactions: Development of the "Trust Perception Scale-HRI," R. Mittu, D. Sofge, A. Wagner, and W. Lawless, Eds., Boston, MA: Springer US, 2016, pp. 191–218. [Online]. Available: https://doi.org/10.1007/978-1-4899-7668-0_10

[3] G. Charalambous, S. Fletcher, and P. Webb, "The Development of a Scale to Evaluate Trust in Industrial Human-robot Collaboration," *International Journal of Social Robotics*, vol. 8, no. 2, pp. 193–209, Apr. 2016, doi: 10.1007/s12369-015-0333-8.

[4] I. Cucciniello, S. Sangiovanni, G. Maggi, and S. Rossi, "Mind Perception in HRI: Exploring Users' Attribution of Mental and Emotional States to Robots with Different Behavioural Styles," *International Journal of Social Robotics*, vol. 15, no. 5, pp. 867–877, 2023, doi: 10.1007/s12369-023-00989-z.

[5] S. Diefenbach, M. Herzog, D. Ullrich, and L. Christoforakos, "Social Robot Personality: A Review and Research Agenda," Springer VS, Wiesbaden, 2023, pp. 217–246. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-658-37641-3_9

[6] R. Maure and B. Bruno, "Autonomy in socially assistive robotics: a systematic review," *Frontiers in Robotics and AI*, vol. 12, p. 1586473, doi: 10.3389/frobt.2025.1586473.

[7] T.-H. Lin, S. Ng, and S. Sebo, "2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)," Aug. 2022, pp. 37–44. doi: 10.1109/RO-MAN53752.2022.9900828.