Responsive Robots

Preliminary Analysis of Trust and Performance Data

Shauna Heron

2025-12-18

> ⚠ Warning
>
> These are preliminary results and analyses. Please do not distribute or cite without permission of the authors.

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.

Introduction

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 . In collaborative settings, trust influences not only subjective evaluations of the robot but also objective outcomes such as task performance, compliance, and engagement. As a result, a growing body of work has focused on measuring trust following human–robot interaction, including the development of standardized instruments designed to capture users' perceptions of robot reliability, predictability, and intent.

Despite this growing literature, much of what is currently known about trust in HRI has been derived from interactions conducted under highly controlled or idealized conditions. In many studies, robot behavior is scripted, simulated, or mediated through human control using Wizard-of-Oz (WoZ) paradigms [1], [2]. While such approaches are valuable for early-stage design and hypothesis generation, these approaches alter interaction dynamics by masking 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.

The present study addresses this gap by evaluating trust following an in-person interaction with a Misty-II robot operating autonomously within predefined behavioral constraints. To this end, participants engaged in solving a mystery 'who-dunnit' style problem with the robot: who took the lab robot 'Atlas' and where is it now? Together the robot and the participant moved through a series of collaborative tasks, the robot managing speech-based interaction, task progression, and affect-responsive behavior, all without human intervention. To this end, two experimental conditions were compared: a control condition in which the

robot followed a neutral, non-proactive interaction policy, and a responsive condition in which the robot was prompted to adapt its behavior based on dialogue, detected user affect and the task itself. Importantly, both conditions were subject to the same sensory and interaction limitations inherent to autonomous operation, including speech recognition variability and response timing constraints.

To this end, we developed an autonomous spoken-language interaction system integrated with a prompted ASR pipeline and the Misty-II robot platform that can engage in natural conversations with users. The system is capable of recognizing speech, managing dialogue, and generating spoken responses as well as physical expression and 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 and Collaborative Structure

Participants interacted with the robot in solving an immersive puzzle game where the robot served as a diegetic "game guide" and collaborative partner. In the game, participants solve a crime mystery by asking the game guide for information to complete tasks and for hints and advice on how to solve puzzles. The game was composed of two sequential tasks designed to elicit interaction with the robot under differing knowledge and dependency conditions T.-H. Lin, S. Ng, and S. Sebo [2] . When T.-H. Lin, S. Ng, and S. Sebo [3] 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. The robot autonomously monitored task progression through the interface and adapted its dialogue accordingly, while all behavioral responses were generated without real-time human intervention.

Task 1: Robot-dependent collaborative reasoning

The first task required participants to identify a suspect from a 6x4 grid 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. In this task, the robot possessed the 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, interpret its responses, 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 examined multiple technical logs presented via the interface to determine the location of the missing 'Atlas' robot. Participants had 3 questions to answer via multiple-choice dropdowns: i) is Atlas still functioning? (yes/no); ii) what building is Atlas is; iii) what floor is Atlas on; iv) what room is Atlas in. 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 the Gemini 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 task-related questions and that its job was to determine Atlas' whereabouts together. The robot could ask the participant questions and vice versa. Importantly, participants could complete the task 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.

Once all answers were submitted, the correct answers were shown to participants, letting them know how they did. At the stage the robot and the participant could briefly debrief on whether they were right or not, and then the task came to the end with the robot thanking them and

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

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 robot operated fully autonomously throughout each session, managing speech recognition, dialogue generation, affect detection, and behavioral responses in real time. No human operator intervened during interactions. All behaviors were constrained to a predefined action space designed to ensure safety and task consistency across participants.

The task interface was adapted from prior work with the same robotic platform in which a graphical interface was used to support Wizard-of-Oz control. In the present study, this interface was deliberately repurposed as a shared workspace for human–robot collaboration rather than as a mechanism for remotely controlling robot behavior. Rather than serving as a control surface, the interface functioned 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.

In the first task, participants worked with the robot to identify a suspect by asking a series of yes/no questions. The robot possessed the ground-truth information necessary to resolve the task but could not explicitly identify the individual directly, making successful completion dependent on effective interaction with the robot.

The second task involved a more complex problem-solving scenario in which participants examined multiple WiFi and other technical logs to determine the location of a missing robot. Unlike the first task, the robot did not possess the correct solution. Participants could choose to work independently or solicit

assistance from the robot, which provided guidance, clarification, and affective support but no definitive answers. The second 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.

During the second task, the robot did not possess task-specific knowledge or access to the correct solution. Instead, it provided assistance by helping participants interpret the structure and purpose of the available technical logs, drawing solely on general knowledge and reasoning capabilities acquired during model training. The robot's responses were conditioned on the interaction context and participant queries, but it did not have access to the log contents beyond what participants explicitly referenced.The robot's dialogue system was explicitly constrained such that it was informed only that participants could view multiple technical logs, without access to the content of those logs or the correct solution to the task.

As a result, the robot's role in the second task was that of a collaborative reasoning partner rather than an authoritative source. Participants could choose whether to engage with the robot's suggestions or pursue independent reasoning strategies.The robot's responses were framed to emphasize collaborative rather than definitive instruction, reducing the risk of misleading participants when uncertainty was present.

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 Table 1).

Table 1: Participant Demographics and Baseline Characteristics by Group

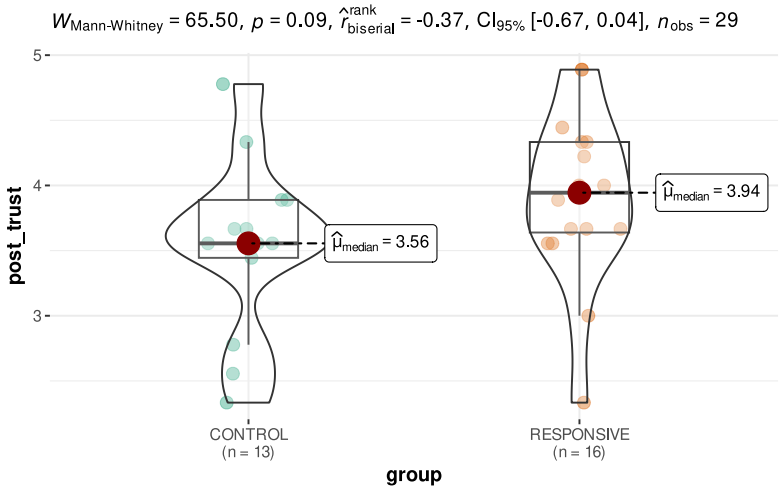
Characteristic	N	CONTROL N = 13[1]	RESPONSIVE N = 16[1]	p-value[2]
Gender	27			0.84
Woman		6 / 13 (46%)	7 / 14 (50%)	
Man		7 / 13 (54%)	7 / 14 (50%)	
Missing		0	2	
Age Group	27			0.35
18-24		5 / 13 (38%)	7 / 14 (50%)	
25-34		4 / 13 (31%)	2 / 14 (14%)	
34-44		1 / 13 (7.7%)	4 / 14 (29%)	
45+		3 / 13 (23%)	1 / 14 (7.1%)	
Missing		0	2	
Program	25			>0.99
Psychology		1 / 13 (7.7%)	1 / 12 (8.3%)	
Engineering		2 / 13 (15%)	1 / 12 (8.3%)	
Computer Science		7 / 13 (54%)	6 / 12 (50%)	
Earth Sciences		0 / 13 (0%)	1 / 12 (8.3%)	
Other		3 / 13 (23%)	3 / 12 (25%)	
Missing		0	4	
Experience w/Robots	29	7 / 13 (54%)	4 / 16 (25%)	0.14
Native English Speaker	29			0.53
Native English		5 / 13 (38%)	8 / 16 (50%)	
Non-Native English		8 / 13 (62%)	8 / 16 (50%)	
NARS Overall	29	38 (8)	38 (7)	0.79
Need for Cognition	29	3.62 (0.78)	3.74 (0.74)	0.55

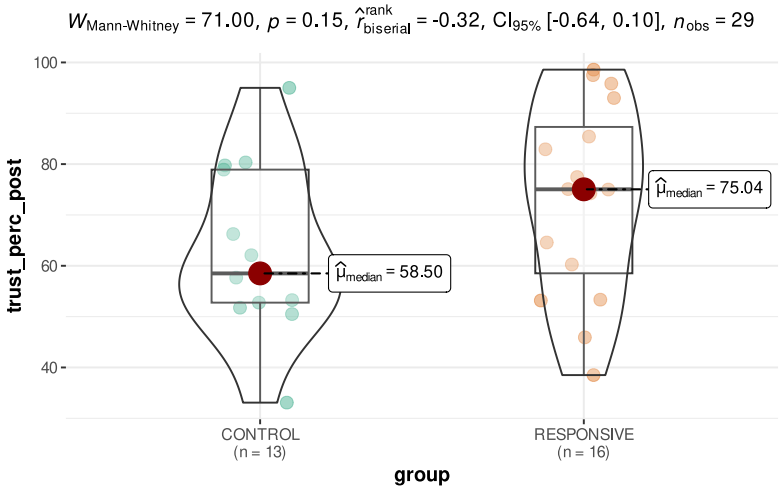
[1] n / N (%); Mean (SD)

[2] Pearson's Chi-squared test; Fisher's exact test; Wilcoxon rank sum test

Post-Interaction Trust Differences

Descriptive comparisons of participant-level post-test scores indicated an approximately 10 point difference in post-test Trust Perception Scale-HRI scores (M ≈ 73 vs M ≈ 63) and a 14 point difference in the Trust in Industrial Human-robot Collaboration scale (M ≈ 47 vs M ≈ 61) between conditions, although these differences did not reach conventional significance under a two-sample t-test (p = 0.09 and p = 0.14 respectively), reflecting limited power (n = 29). Although the second trust instrument was administered and scored in its original 5-point Likert format, for ease of interpretation it was linearly rescaled to the first scale's 0–100 metric for interpretability and comparability.





Hierarchical Bayesian models were next employed to account for item-level structure yielding smaller effect sizes but consistent estimates (≈7–8 points) and a high posterior probability of a positive effect. Notably, treating item-level responses as independent observations (i.e., ignoring non-independance) substantially inflated apparent precision, underscoring that the observed uncertainty is primarily a function of sample size rather than absence of an effect. A larger sample would allow more precise estimation of the effect magnitude.

To explore further, Bayesian hierarchical models indicated higher post-interaction trust scores in the responsive robot condition across both trust-related scales (posterior mean differences ≈ 7–8 points on a 0–100 scale). Although 95% credible intervals overlapped zero, the posterior probability that the responsive condition increased trust was high for both measures (≈0.95), suggesting a robust directional effect alongside substantial individual variability.Sensitivity analyses using substantially wider priors yielded nearly identical posterior estimates for the group effect, indicating that results were not driven by prior specification. In addition to directional effects, the posterior probability that the responsive condition increased trust by at least five points was approximately 0.70, suggesting a moderate likelihood of a practically meaningful effect despite substantial individual variability.

Trust subscale patterns

Examination of trust subscales suggested that group differences were most pronounced for affective components of trust (e.g., trust feelings), whereas differences in perceived reliability were smaller. This pattern aligns with correlations showing that baseline negative attitudes toward robots were more strongly associated with affective trust than with reliability judgments.

Interaction dynamics and task performance

Characteristic	N	CONTROL N = 13[1]	RESPONSIVE N = 16[1]	p-value[2]
post_trust	29	47 (26)	61 (26)	0.094
post_trust_reliability	29	46 (24)	62 (20)	0.11
post_trust_perception	29	49 (26)	57 (25)	0.32
post_trust_feelings	29	59 (33)	72 (29)	0.26
Post-Task Trust Perception	29	63 (17)	73 (19)	0.16
Suspect ID Accuracy	29	4 / 13 (31%)	10 / 16 (63%)	0.089
Status Accuracy	29	11 / 13 (85%)	10 / 16 (63%)	0.24
building_correct	29	10 / 13 (77%)	12 / 16 (75%)	>0.99
floor_correct	29	10 / 13 (77%)	13 / 16 (81%)	>0.99
zone_correct	29	6 / 13 (46%)	4 / 16 (25%)	0.27
Total Task Accuracy	29	3.15 (0.99)	3.06 (1.34)	0.98
Overall Task Accuracy	29	0.63 (0.20)	0.61 (0.27)	0.98
Dialogue Turns	29	32 (10)	36 (11)	0.95

[1] Mean (SD); n / N (%)

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test; Pearson's Chi-squared test; Fisher's exact test

Characteristic	N	CONTROL N = 13[1]	RESPONSIVE N = 16[1]	p-value[2]
Avg Task Duration (mins)	29	12.84 (4.02)	16.81 (6.38)	**0.050**
Avg Response Time (ms)	29	15.1 (4.1)	17.2 (2.4)	**0.006**
Silent Periods	29	5.15 (2.27)	5.31 (2.82)	0.88
Engaged Responses	29	1.92 (2.25)	3.50 (1.83)	**0.020**
Frustrated Responses	29	0.54 (0.66)	0.88 (1.15)	0.56
n_neg	29	1.00 (0.91)	0.94 (1.18)	0.63

[1] Mean (SD); n / N (%)

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test; Pearson's Chi-squared test; Fisher's exact test

Task performance

Objective task accuracy did not differ between conditions across any task-level measures except suspect accuracy (robot dependendant task), indicating that increased trust was only attributable to improved task success when interaction was necessary to complete accurately.

Despite similar task accuracy, interactions in the responsive condition were characterized by longer durations, slower response times, and a higher number of AI-detected engaged responses. These findings suggest that responsiveness altered the interaction dynamics and affective tone rather than task outcomes.

Individual differences and correlational patterns

As expected, we find that higher Need for Cognition (NFC) scores are negatively associated with Negative Attitudes Towards Robots (NARS), indicating that individuals who enjoy effortful thinking tend to have more positive attitudes towards robots. This relationship is consistent with prior literature suggesting that cognitive engagement is associated with openness to new technologies. In terms of NARS subscales, NFC was negatively correlated with all three subscales, but significantly so only in the domain of Situations of Interaction with Robots. This suggests that individuals with higher NFC are less likely to hold negative attitudes across various dimensions of robot interaction but especially around direct interaction with robots.

–> how to talk about post-interaction correlations w/pre-interaction measures Several behavioral and task-level measures were correlated with post-interaction trust, consistent with the interpretation that trust judgments were shaped by interaction quality; these variables were not included as covariates in primary models to avoid conditioning on potential mediators.

Baseline negative attitudes toward robots were negatively correlated with post-interaction trust, with the strongest associations observed for affective trust subscales. In contrast, objective task performance was selectively associated with perceived reliability. Need for cognition was negatively correlated with negative robot attitudes and interaction-level negative affect, suggesting that individual differences contributed to variability in trust responses.

	nars_pre	*nfc_pre*	*nars_social_influence_robots*	*nars_emotion_robots*	*nars_interaction_robots*
nars_pre					
nfc_pre	−0.482**				
nars_social_influence_robots	0.848***	−0.227			
nars_emotion_robots	0.616***	−0.224	0.426*		
nars_interaction_robots	0.869***	−0.516**	0.680***	0.410*	

Computed correlation used spearman-method with listwise-deletion.

Figure 2

Model robustness and predictive checks

Sensitivity analyses using alternative prior specifications yielded substantively similar estimates, and leave-one-out cross-validation indicated comparable predictive performance between models with and without the group effect.

Discussion

Descriptive comparisons of post-interaction measures indicated that participants in the responsive condition reported consistently higher trust across all trust measures, with differences ranging from approximately 8 to 16 points on a 0–100 scale, although uncertainty remained high given the small sample. Notably, the responsive condition did not differ from control in objective task accuracy, suggesting that increased trust was not driven by improved task success. Instead, responsive interactions were characterized by longer durations, slower response times, and a higher number of AI-detected engaged responses, indicating a shift in interaction dynamics rather than performance.

Baseline negative attitudes toward robots were most strongly associated with affective components of trust rather than perceptions of reliability, suggesting that pre-existing attitudes primarily shape emotional responses to interaction rather than judgments of system competence. Conversely, objective task performance was selectively associated with perceived reliability, indicating that participants distinguished between affective and functional aspects of trust.

Future work with larger samples could formally test mediation pathways linking robot responsiveness, interaction fluency, affective responses, and trust judgments, as well as moderation by baseline attitudes toward robots and need for cognition.

Participants in the responsive condition also exhibited higher levels of AI-detected engagement during interaction, as indexed by a greater number of responses classified as positive affect (t-test result). This suggests that responsive behaviors altered the affective tone of the interaction itself.

Bibliography

[1] 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.

[2] 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.

[3] 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.