# Responsive Robotics to Increase 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

2026-01-03

**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.

## Introduction

As automation expands across safety-critical domains such as manufacturing, mining, and healthcare, robotic systems are increasingly expected to operate alongside humans rather than in isolation [1], [2], [3], [4]. In these collaborative settings, successful deployment depends not only on technical performance and safety guarantees, but on whether human users are willing to rely on, communicate with, and coordinate their actions around systems driven by artificial intelligence (AI) [5], [6]. Trust has therefore emerged as a central determinant of adoption and effective use in human–robot collaboration (HRC) [5], [7]. Insufficient trust can lead to disuse or rejection of automation, while excessive trust risks overreliance—particularly in environments characterized by uncertainty or incomplete information [8].

A substantial body of human–robot interaction (HRI) research has examined how robot behaviour shapes user trust, perceived reliability, and cooperation across industrial and social contexts [9], [10]. Trust is commonly conceptualized as a multidimensional construct encompassing cognitive evaluations of competence and reliability, affective responses to the interaction partner, and behavioural willingness to rely on the system under conditions of risk or uncertainty [8], [11], [12]. Despite this multidimensional framing, empirical studies have predominantly operationalized trust using post-interaction self-report questionnaires, often collected following short, highly controlled interactions.

Importantly, much of the existing HRI trust literature relies on scripted behaviours, simulated environments, or Wizard-of-Oz paradigms in which a human operator covertly manages the robot's behaviour. While these approaches are valuable for isolating specific design factors, they obscure the interaction breakdowns and system imperfections that characterize real-world autonomous robots [5]. In deployed systems, limitations such as speech recognition errors, delayed responses, misinterpretations of user intent, and incomplete affect sensing are not peripheral issues but defining features of interaction. These failures are likely to play a decisive role in shaping trust and collaboration, yet remain underrepresented in empirical evaluations.

One proposed mechanism for supporting trust in HRI is responsiveness: the extent to which a robot adapts its behaviour based on user state and interaction context [9], [10]. Responsive robots may adjust dialogue, timing, or support strategies in response to inferred cues such as confusion, frustration, or disengagement, and prior work suggests that such adaptive behaviour can enhance perceived social intelligence and trustworthiness in dialogue-driven tasks [13]. However, most evidence for these effects comes from simulated or semi-autonomous systems, leaving open questions about how responsiveness operates when implemented in fully autonomous, in-person interactions subject to real-time constraints and failure [5].

From an engineering perspective, responsiveness represents an interaction policy rather than a superficial social cue [9]. Proactive assistance based on interaction context differs fundamentally from reactive, request-based behaviour, particularly in fully autonomous systems—for example, offering clarification or encouragement when confusion or hesitation is inferred, rather than waiting for an explicit request for help [13]. Implementing such policies requires robots to manage spoken-language dialogue, track interaction state over time, and coordinate verbal and nonverbal responses in real time, all while operating under noise, latency, and sensing uncertainty [5].

The present work addresses these gaps through a pilot study examining trust and collaboration during in-person interaction with a fully autonomous social robot. Participants collaborated with one of two versions of the same robot platform during a dialogue-driven puzzle task requiring shared problem solving. In both conditions, all interaction management—including speech recognition, dialogue state tracking, task progression, and response generation—was handled and logged autonomously by the robot without human intervention. In the responsive condition, the robot employed a proactive interaction policy, adapting its assistance based on conversational cues and inferred user affect. In the neutral condition, the robot followed a reactive policy, providing general guidance but assistance only when explicitly requested.

This pilot study had three primary objectives: (1) to design and evaluate the feasibility of an autonomous spoken-language interaction system with affect-responsive behaviour on a mobile robot platform; (2) to assess whether interaction policy influences post-interaction trust and collaborative experience under realistic autonomous conditions; and (3) to explore how behavioural and interaction-level indicators align with subjective trust evaluations. Rather than optimizing for flawless interaction, the system was intentionally designed to reflect the capabilities and limitations of contemporary social robots, allowing interaction breakdowns to surface naturally.

By combining post-interaction trust measures with task-level and behavioural observations, this study aims to contribute empirical evidence on how trust in human–robot collaboration emerges and is enacted during fully autonomous interaction. The findings are intended to inform the design of a larger subsequent study by evaluating feasibility and identifying technical, interactional, and methodological challenges that must be addressed when evaluating affect-responsive robots in real-world contexts.

# Methods

## Experimental Design and Conditions

This study employed a between-subjects experimental design to examine how robot interaction policy influences trust and collaboration during fully autonomous, in-person human–robot interaction. The sole experimental factor was the robot's interaction policy, with participants randomly assigned to interact with either a responsive or neutral version of the same robot system.

Throughout this paper, references to *the robot* denote the fully autonomous interactive system comprising the Misty-II hardware platform and its onboard software stack, with all interaction decisions generated without human intervention, including spoken-language processing, dialogue management, and the interaction policy governing verbal and nonverbal behaviour. Additional details of the system architecture are provided in Appendix A.

Participants interacted with a Misty-II social robot in a shared physical workspace that included a participant-facing computer interface [14]. The interface was used to display task materials, collect participant inputs, and manage task progression (see Figure 2). Importantly, the interface did not function as a control mechanism for the robot. Instead, the robot autonomously monitored task state and participant inputs via the interface and managed dialogue and behaviour accordingly, without real-time human intervention.

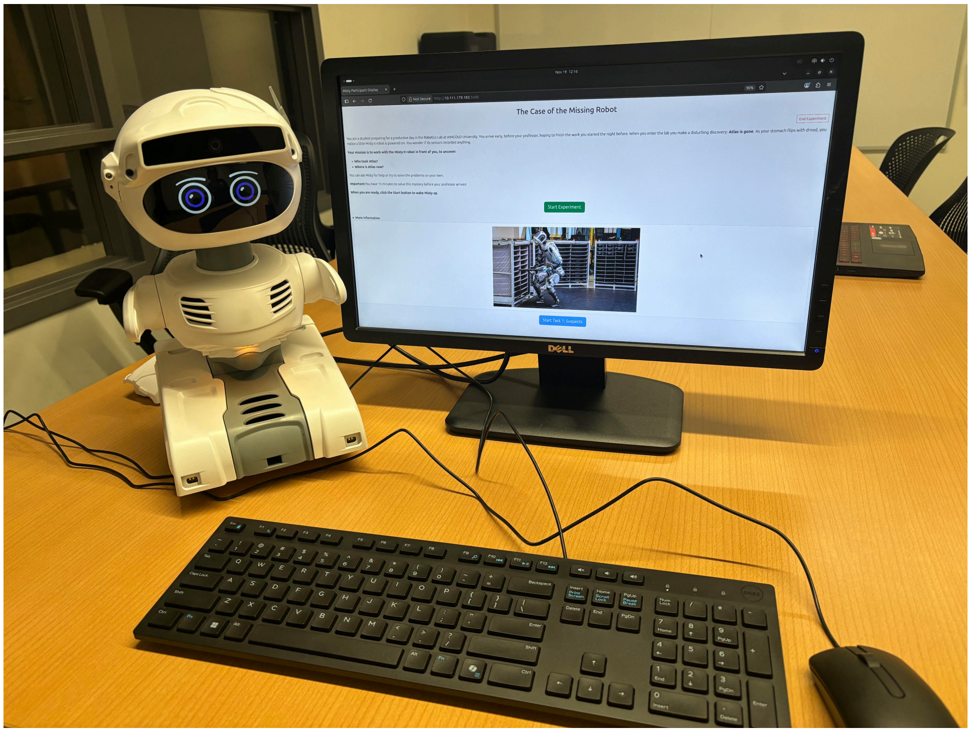
### Interaction Policies

- **Responsive condition (experimental):**
  The robot employed a proactive, affect-adaptive interaction policy. Robot responses were modulated based on inferred participant affect, dialogue context, and task demands, resulting in unsolicited encouragement, clarification, and engagement-oriented behaviours when appropriate.
- **Control condition (baseline):**
  The robot employed a neutral, reactive interaction policy. General information and guidance were were provided but additional help only when explicitly requested by the participant and without affect-based adaptation or proactive support beyond a check-in when participant was silent for more than 1 minute.

Both conditions used identical hardware, software infrastructure, sensing capabilities, and task logic. The only difference between conditions was the robot's interaction policy.

### Collaborative Task Design

Participants completed an immersive, narrative-driven puzzle game consisting of five sequential stages and two timed reasoning tasks. The game context positioned participants as investigators searching for a missing robot colleague, with the robot serving as a diegetic guide and collaborative partner. The overall interaction lasted approximately 25 minutes.



You are a student preparing for a productive day in the Robotics Lab at IAMGOLD University. You arrive early, before your professor, hoping to finish the work you started the night before. When you enter the lab you make a disturbing discovery: **Atlas is gone**. As your stomach flips with dread, you notice a little Misty-II robot is powered on. You wonder if its sensors recorded anything.

**Your mission is to work with the Misty-II robot in front of you, to uncover:**

- **Who took Atlas?**
- **Where is Atlas now?**

You can ask Misty for help or try to solve the problems on your own.

**Important:**You have 15 minutes to solve this mystery before your professor arrives!

**When you are ready, click the Start button to wake Misty up.**

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.

The task structure was designed to elicit collaboration under two distinct dependency conditions: (1) enforced collaboration, where the robot was required to complete the task, and (2) optional collaboration, where participants could choose whether to engage the robot.

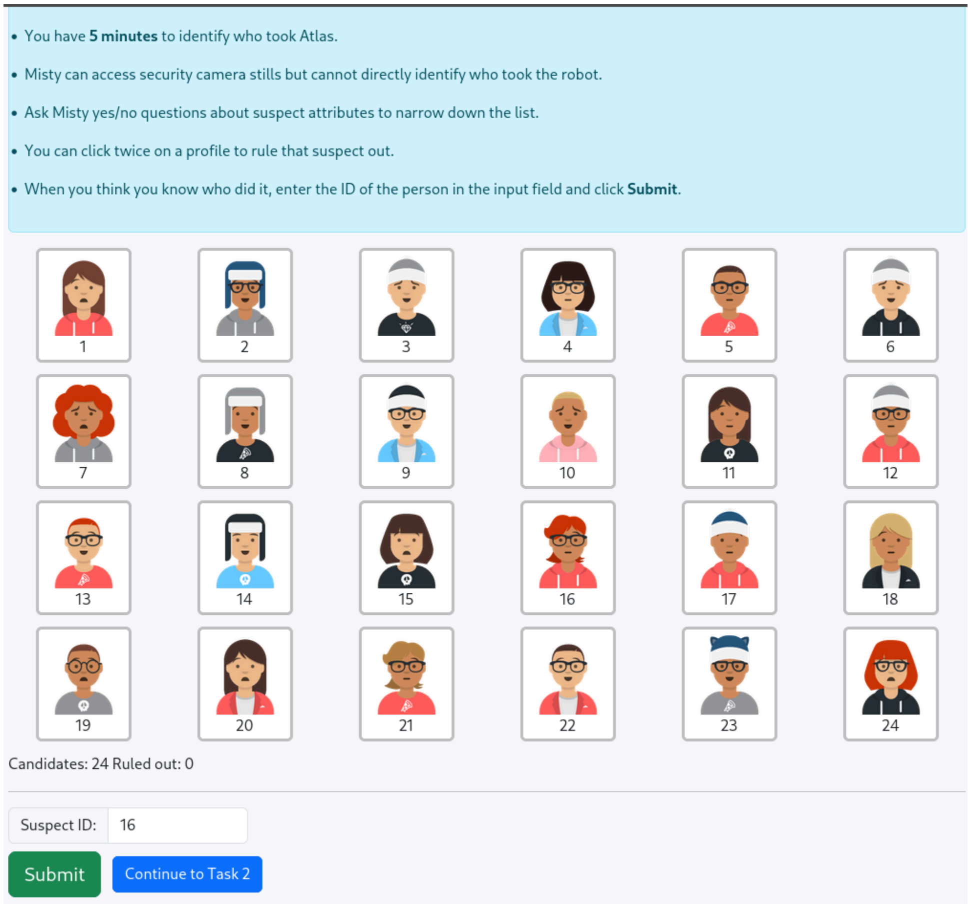
### Stage Overview

1. **Greeting:** The robot introduced itself and engaged in brief rapport-building dialogue.
2. **Mission Brief:** The robot explained the narrative context and overall objectives.
3. **Task 1:** Robot-dependent collaborative reasoning task.
4. **Task 2:** Open-ended problem solving with optional robot support.
5. **Wrap-up:** The robot provided closing feedback and concluded the interaction.

Participants advanced between stages using the interface, either at the robot's prompting or at their own discretion. All spoken dialogue and interaction events were logged automatically.

### Task 1: Robot-Dependent Collaborative Reasoning

In Task 1, participants were asked to identify a perpetrator from a 6 × 4 grid of 24 'suspects' by asking the robot a series of yes/no questions about the suspect's features (e.g., "was the suspect wearing a hat?"). The grid was displayed on the interface, while questions were posed verbally.



- You have **5 minutes** to identify who took Atlas.
- Misty can access security camera stills but cannot directly identify who took the robot.
- Ask Misty yes/no questions about suspect attributes to narrow down the list.
- You can click twice on a profile to rule that suspect out.
- When you think you know who did it, enter the ID of the person in the input field and click **Submit**.

Candidates: 24 Ruled out: 0

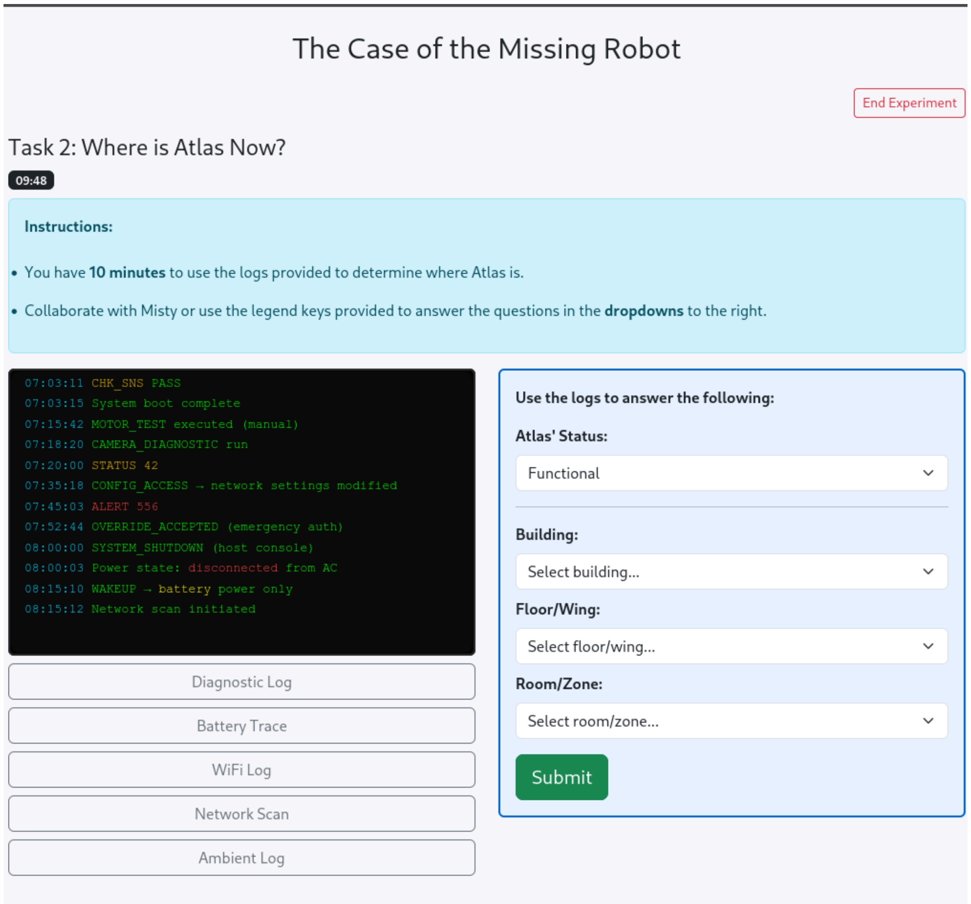
Suspect ID: 16

Figure 2: *Task 1 interface including the 6 × 4 grid of 24 candidates. Participants could track those eliminated by clicking on subjects which would grey them out. A box was provided to input their final answer and a button included to move to the Task 2 interface.*

The robot possessed ground-truth information necessary to answer each question correctly. Successful task completion was therefore dependent on interaction with the robot, creating a forced collaborative dynamic. Participants were required to coordinate questioning strategies with the robot to narrow down the suspect within a five-minute time limit. The structured nature of the task ensured consistent interaction demands across participants and conditions.

### Task 2: Open-Ended Collaborative Problem Solving

Task 2 involved a more open-ended reasoning challenge. Participants were presented with multiple technical logs through a simulated terminal interface that could be used to infer the location of the missing robot.



The Case of the Missing Robot

Task 2: Where is Atlas Now?

09:48

**Instructions:**

- You have **10 minutes** to use the logs provided to determine where Atlas is.
- Collaborate with Misty or use the legend keys provided to answer the questions in the **dropdowns** to the right.

```
07:03:11 CHK_SNS PASS
07:03:15 System boot complete
07:15:42 MOTOR_TEST executed (manual)
07:18:20 CAMERA_DIAGNOSTIC run
07:20:00 STATUS 42
07:35:18 CONFIG_ACCESS → network settings modified
07:45:03 ALERT 556
07:52:44 OVERRIDE_ACCEPTED (emergency auth)
08:00:00 SYSTEM_SHUTDOWN (host console)
08:00:03 Power state: disconnected from AC
08:15:10 WAKEUP → battery power only
08:15:12 Network scan initiated
```

Use the logs to answer the following:

Atlas' Status: Functional

Building: Select building...

Floor/Wing: Select floor/wing...

Room/Zone: Select room/zone...

Figure 3: The task 2 interface presented multiple technical logs through a simulated terminal interface that could be used to determine the location of the missing robot.

Unlike Task 1, the robot did not have access to ground-truth information or the contents of the logs. The robot's assistance was limited to general reasoning support derived from its language model, such as explaining how to interpret log formats, suggesting problem-solving strategies, or prompting participants to reflect on inconsistencies.

Participants could complete this task independently or solicit assistance from the robot at their discretion [15]. This design allowed collaboration to emerge voluntarily rather than being enforced by task structure, positioning the robot as a collaborative partner rather than an authoritative source.

### Study Protocol

Participants signed up for the study and completed a pre-session questionnaire before their in-person session via Qualtrics. The pre-session questionnaire colleced basic demographics information and assessed baseline characteristics, including the Negative Attitudes Toward Robots Scale (NARS) and the short form of the Need for Cognition scale (NFC-s). These measures were used to capture individual differences that may moderate responses to robot interaction.

In-person sessions were conducted in a quiet, private room at Laurentian University between November and December 2025. Prior to each session, the robot's interaction policy was configured to the assigned experimental condition.

Upon arrival, participants were greeted by the researcher, provided with a brief overview of the session, and given instructions for effective communication with the robot, including waiting for a visual indicator before speaking. Once participants indicated readiness, the researcher exited the room, leaving the participant and robot to complete the interaction without human presence or observation. Participants initiated the interaction by clicking a start button on the interface and were informed that they could terminate the session at any time without penalty.

Following task completion, participants completed a 21-item post-interaction questionnaire assessing trust. Participants then engaged in a brief debrief with the researcher and were awarded a $15 gift card. Total session duration averaged approximately 30 minutes.

## Measures

A combination of self-report and objective measures was used to assess trust, engagement, and task performance.

### Self-Report Measures

Trust was assessed using two established self-report instruments commonly used in human–robot interaction research: the Trust Perception Scale–HRI (TPS-HRI) and the Trust in Industrial Human–Robot Collaboration scale (TI-HRC) [16], [17]. Both measures were adapted to reflect the specific task context and interaction modality of the present study. 9 items were retained from the TI-HRC and 12 items from the TPS-HRI. Item wording was modified to reference the robot's behaviour during a dialogue-driven collaborative task, and response formats were adjusted to ensure interpretability for participants without prior robotics experience.

Together, these instruments capture complementary dimensions of trust, including perceived reliability, task competence, and affective comfort. However, they differ in their conceptual emphasis: the TPS-HRI primarily operationalizes trust as a reflective judgement of system performance (i.e., "What percent of the time was the robot reliable"), whereas the TI-HRC scale emphasizes trust as an experienced, embodied response arising during interaction (i.e., "The way the robot moved made me feel uneasy"). Despite this complementarity, both measures rely on retrospective self-report and may be insensitive to moment-to-moment trust dynamics as collaboration unfolds. For this reason, questionnaire data were interpreted alongside behavioural and interaction-level measures.

Participants completed a pre-session questionnaire assessing baseline characteristics, including the Negative Attitudes Toward Robots Scale (NARS) and the short form of the Need for Cognition scale (NFC-s). These measures were used to capture individual differences that may moderate responses to robot interaction.

### Objective and behavioural Measures

Objective task metrics included task completion, task accuracy, time to completion, and the number of assistance requests made to the robot. behavioural engagement metrics were derived from interaction logs and manually coded dialogue transcripts, including number of dialogue turns, frequency of communication breakdowns, response timing, and task-relevant robot contributions.

## Participants, Communication Viability, and Analytic Strategy

A total of 29 participants were recruited from the Laurentian University community via word of mouth and the SONA recruitment system. Eligibility criteria included being 18 years or older, fluent in spoken and written English, and having normal or corrected-to-normal hearing and vision. Participants received a $15 gift card as compensation for their time. All procedures were approved by the Laurentian University Research Ethics Board (REB #6021966).

Although English fluency was an eligibility requirement, in-person observation during data collection indicated meaningful variability in participants' functional spoken-language proficiency. The researcher therefore recorded observed English proficiency for each session in anticipation of potential speech-based system limitations. Subsequent post-hoc review of interaction transcripts and system logs revealed that a subset of sessions exhibited severe and sustained communication failure. In these cases, automatic speech recognition (ASR) output was largely unintelligible or fragmented, preventing the robot from extracting sufficient linguistic content to maintain dialogue, respond meaningfully to participant queries, or support task progression. Interaction frequently stalled, participant input went unanswered or was misinterpreted, and collaborative problem-solving was not feasible. These sessions reflected a breakdown of language-mediated interaction, rendering the experimental manipulation inoperative.

Because the study relied fundamentally on spoken-language collaboration, sessions exhibiting persistent communication failure were classified as protocol non-adherence and excluded from task-level analyses (n = 6). Exclusion decisions were based solely on communication viability and interaction mechanics, not on task outcomes or trust measures.

Table 1: Participant Demographics and Baseline Characteristics by Group

| | | **CONTROL** | **RESPONSIVE** | |
|---|---|---|---|---|
| **Characteristic** | **N** | N = 13[1] | N = 16[1] | **p-value**[2] |
| **Gender** | 27 | | | 0.84 |
| Woman | | 6 / 13 (46%) | 7 / 14 (50%) | |
| Man | | 7 / 13 (54%) | 7 / 14 (50%) | |
| **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%) | |
| **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%) | |
| **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 |
| **Dialogue Viability** | 29 | | | 0.63 |
| exclude | | 3 / 13 (23%) | 2 / 16 (13%) | |
| include | | 10 / 13 (77%) | 14 / 16 (88%) | |

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

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

To ensure transparency and assess the impact of communication-based exclusions, analyses were conducted in three stages. First, an eligible-sample analysis (excluding non-viable sessions) served as the primary analysis, reflecting interactions in which the spoken-language protocol and experimental manipulation operated as intended. Second, a full-sample analysis including all participants was conducted as a sensitivity analysis to evaluate robustness to communication failures and protocol deviations. Third, a mechanism-focused analysis compared included and excluded sessions on interaction-process metrics (e.g., ASR failure rates, dialogue turn completion, task abandonment) to characterize how severe communication breakdown alters interaction dynamics.

While full-sample analyses are informative as robustness checks, trust measures obtained from sessions with complete communication breakdown are not interpreted as valid estimates of human–robot trust under functional interaction. In these cases, the robot was unable to sustain dialogue or collaborative behaviour, precluding meaningful evaluation of reliability, competence, or collaborative intent.

Across analyses, participants in the responsive and control conditions were comparable with respect to demographic characteristics, prior experience with robots, and baseline attitudes toward robots, including Negative Attitudes Toward Robots (NARS) and Need for Cognition scores (see Table 1) [18]. These patterns were consistent across both eligible and full samples, indicating successful random assignment.

# Results

## Communication Viability and Analytic Samples

Prior to hypothesis testing, interaction sessions were classified based on communication viability using a dialogue-level metric derived from system logs and manual coding. Specifically, the proportion of dialogue turns affected by speech-recognition failure or fragmented utterances was computed for each session. Sessions in which more than 60% of dialogue turns (half of all turns were dependent on human speech) were characterized by communication breakdown were classified as non-viable (n=5). This criterion closely matched sessions independently flagged during administration and reflects cases in which sustained spoken-language interaction was not possible. Of the 29 completed sessions, 5 were classified as non-viable due to severe and persistent communication failure resulting in unintelligble sentence fragments.

Because the experimental manipulation relied on language-mediated collaboration, analyses were conducted using three complementary approaches: (1) a primary eligible-sample analysis excluding non-viable sessions, (2) a full-sample sensitivity analysis including all sessions, and (3) a mechanism-focused analysis examining how communication breakdown altered interaction dynamics.

Unless otherwise noted, inferential results reported below refer to the eligible sample.

## Primary Analysis: Eligible Sample

### Descriptive Outcomes

Descriptive comparisons of post-interaction trust measures indicated higher trust ratings in the responsive condition relative to the control condition across both trust scales (see Table 2).

Average post-interaction scores on the TI-HRC differed by approximately 26 points (Likert 1-5 converted to 0-100 scale for easier comparison across scales). While differences in TPS-HRI scores were approximately 15 points higher in the responsive condition compared to the control. Scores on the Behavioural summaries further indicated differences in dialogue patterns and robot assistance behaviours consistent with the intended interaction policies.

Importantly objective task accuracy did not differ between conditions across any task-level measures. This suggests that observed differences in trust were not driven by differential task success.

Despite similar task accuracy, interactions in the responsive condition were expectedly characterized by longer durations (more dialogue), slower robot response times (more dialogue), 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.

Table 2: Post-Interaction Raw Outcome Measures by Group

| Characteristic | CONTROL<br>N = 10[1] | RESPONSIVE<br>N = 14[1] | p-value[2] |
|---|---|---|---|
| Trust in Industrial HRI Collaboration | 39 (22) | 67 (21) | **0.004** |
| Subscales | | | |
| Reliability subscale | 40 (24) | 65 (18) | **0.012** |
| Trust Perception subscale | 42 (23) | 60 (22) | 0.075 |
| Affective Trust subscale | 50 (31) | 79 (22) | **0.018** |
| Trust Perception Scale–HRI | 59 (17) | 77 (18) | **0.022** |
| Overall Task Accuracy | 0.60 (0.21) | 0.66 (0.23) | 0.47 |
| Objective Measures | | | |
| Dialogue Turns | 34 (9) | 33 (5) | 0.45 |
| Avg Session Duration (mins) | 13.24 (3.06) | 15.26 (2.12) | 0.084 |
| Avg Robot Response Time (ms) | 14.37 (3.76) | 17.24 (2.52) | **<0.001** |
| Silent Periods | 5.60 (1.96) | 4.71 (2.05) | 0.29 |
| Engaged Responses | 2.00 (2.21) | 3.50 (1.95) | **0.040** |
| Frustrated Responses | 0.60 (0.70) | 0.93 (1.21) | 0.68 |

[1] Mean (SD)

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test

### Hierarchical Models of Post-Interaction Trust

To evaluate condition effects and to control for pre-test covariates on post-interaction trust, linear mixed-effects models were fitted separately for each trust outcome. All models included interaction policy (RESPONSIVE vs. CONTROL) as the primary fixed effect, along with baseline negative attitudes toward robots (NARS) and native English fluency as baseline covariates. Random intercepts for session were included in all models to account for repeated measurement at the participant level.

Model building proceeded by comparing a baseline model containing interaction policy alone against models incorporating theoretically motivated covariates. Adding baseline negative attitudes toward robots significantly improved model fit ($\chi^2 = 4.82$, p = .028), whereas prior experience with robots did not. While Native English fluency did not significantly improve model fit it was retained as a covariate due to its relevance for spoken-language interaction viability.

### Trust in Industrial Human–Robot Collaboration

For this outcome, inclusion of random intercepts for individual trust items significantly improved model fit, indicating meaningful item-level variability beyond session-level differences. The final model: robot_trust_post ~ policy + nars_pre_c + native_english + (1 | session_id) + (1 | trust_items).

In the final model predicting Trust in Industrial Human–Robot Collaboration, participants who interacted with the responsive robot reported significantly higher post-interaction trust than those in the control condition (β = 16.28, SE = 5.14, t = 3.17, p = .005). Higher baseline negative attitudes toward robots were associated with lower trust scores (β = −7.43, SE = 2.81, p = .016). Native English fluency was not significantly associated with trust, although the estimated effect was negative.

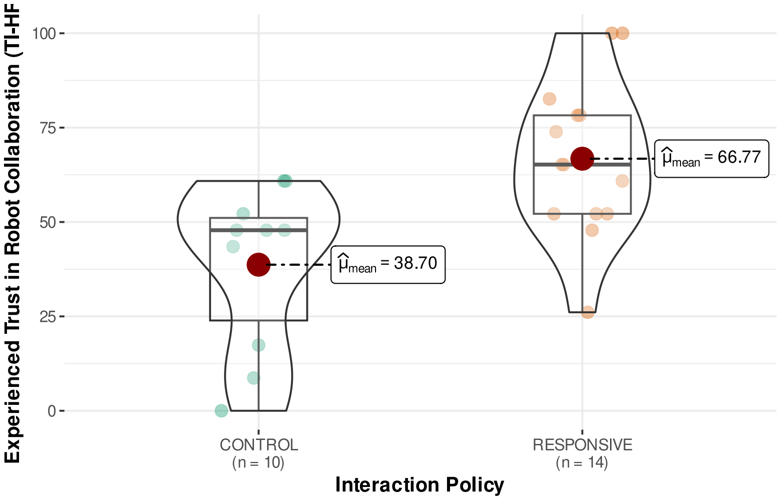


Figure 4: Distribution of Trust in Industrial Human–Robot Collaboration scores by interaction policy. Points represent individual observations; violins depict score distributions. Red points indicate group means with 95% confidence intervals. Statistical comparisons are reported in the Results section.

### Trust Perception Scale–HRI

For the Trust Perception Scale–HRI, a comparable mixed-effects model was fitted using the same fixed effects structure. In this model, interaction with the responsive robot was associated with higher post-interaction trust scores (β = 14.17, SE = 6.5, t = 2.00, p = 0.046). Effects of baseline negative attitudes toward robots and native English fluency followed a similar directional pattern but did not reliably differ from zero.

In contrast to the collaboration trust scale, inclusion of random intercepts for individual trust items did not improve model fit for the Trust Perception Scale–HRI and was therefore omitted. This divergence likely reflects differences in scale format and response interface: the Trust

Perception scale was administered using a continuous slider input, whereas the Trust in Industrial Human–Robot Collaboration scale employed discrete Likert-style response options.

Informal observation during administration and post-hoc inspection of item-level variance suggest that the slider-based interface, administered via a touchpad, may have reduced response precision relative to discrete response formats. While this likely attenuated item-level variability, the Trust Perception Scale–HRI nevertheless captured meaningful between-condition differences at the aggregate level.

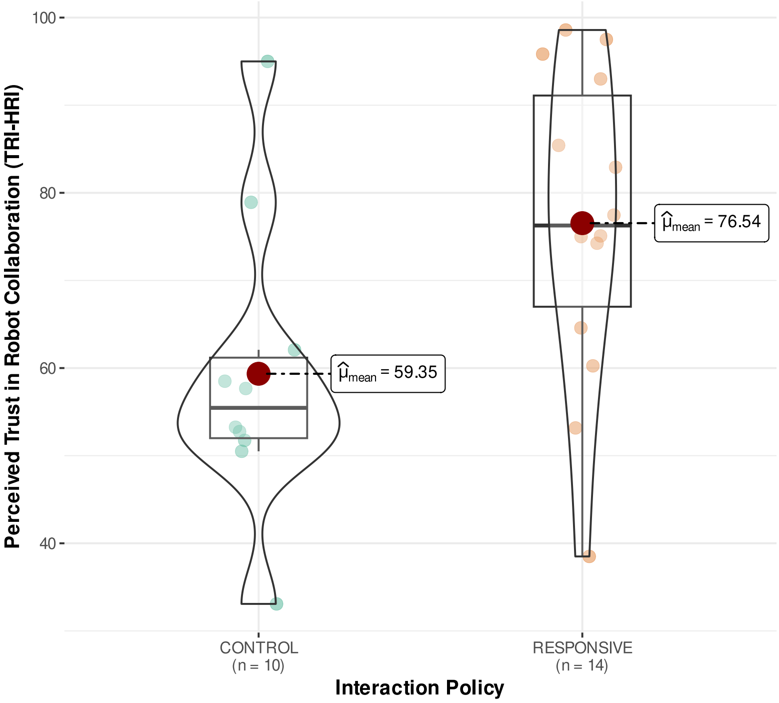


Figure 5: Distribution of Trust Perception in HRI by interaction policy. Points represent individual observations; violins depict score distributions. Red points indicate group means with 95% confidence intervals. Statistical comparisons are reported in the Results section.

Together, these models indicate that robot responsiveness had a consistent positive effect on post-interaction trust, with effect magnitude and measurement sensitivity varying by trust dimension and scale format.

### Bayesian analysis

To complement frequentist mixed-effects models, Bayesian hierarchical models were fitted separately for each trust outcome using weakly informative priors. Models predicted post-interaction trust as a function of interaction policy (RESPONSIVE vs. CONTROL), with random intercepts for session and trust scale items to account for repeated measurement and item-level variability.

Trust in Industrial Human–Robot Collaboration was analysed using Bayesian linear mixed-effects models with random intercepts for session and trust item to account for repeated measurement and item-level heterogeneity. Models included interaction policy (responsive vs. control), baseline negative attitudes toward robots (NARS), and native English fluency as fixed effects. Separate models were fit for the eligible sample (sessions with viable spoken-language interaction) and the full sample (including sessions with severe communication breakdown) to assess sensitivity to protocol adherence.

Eligible sample

In the eligible sample, posterior estimates indicated a clear and substantial effect of interaction policy on experienced trust. Participants who interacted with the responsive robot exhibited higher TI-HRC scores than those in the control condition (posterior mean $\beta = 12.60$, 95% credible interval [2.70, 22.03]). The posterior probability that the effect was positive was high ($P(\beta > 0) = 0.99$), with a 0.71 probability that the effect exceeded a moderate magnitude ($\beta > 10$).

Baseline negative attitudes toward robots were weakly negatively associated with trust, although the credible interval included zero. In contrast, native English fluency showed a credible negative association with TI-HRC scores ($\beta = -10.43$, 95% CI [−20.29, −0.35]), indicating lower experienced trust among non-native English speakers even in interactions where dialogue remained viable.

Random-effects estimates indicated substantial variability at the session level, while item-level variance was comparatively small, suggesting that affective and perceptual trust items formed a relatively coherent construct under functional interaction conditions.

Full sample (sensitivity analysis)

When the same model was fit to the full sample, including sessions with severe and sustained communication breakdown, posterior estimates for the interaction policy effect were markedly attenuated. Although the posterior mean for the responsive condition remained positive ($\beta \approx 7$), the credible interval spanned zero, and the probability of a moderate or larger effect dropped substantially ($P(\beta > 10) < 0.30$).

Residual variance increased in the full sample, indicating greater unexplained variability in trust scores when interactions with non-functional dialogue were included. This pattern suggests that trust ratings obtained under conditions of complete communication failure do not reflect graded variation in experienced trust, but rather a distinct interaction regime in which collaborative behaviour could not be sustained.

Across both samples, posterior estimates for baseline negative attitudes toward robots remained consistently negative, while the effect of native English fluency was reduced and no longer credibly different from zero in the full sample.

Summary

Taken together, these results indicate that responsive interaction policies are associated with higher experienced trust under conditions where spoken-language collaboration is viable. Including sessions characterized by complete communication breakdown substantially increases uncertainty and attenuates estimated policy effects, consistent with the interpretation that trust judgments in these sessions are not meaningful estimates of trust under functional human–robot collaboration. For this reason, eligible-sample analyses are treated as primary, with full-sample results reported as sensitivity analyses reflecting real-world failure conditions rather than alternative estimates of trust.

### Trust Perception Scale–HRI (TPS-HRI)

Task-oriented trust was analysed using Bayesian linear mixed-effects models with the same fixed and random effects structure used for the TI-HRC outcome. Models included interaction policy, baseline negative attitudes toward robots (NARS), and native English fluency as fixed effects, with random intercepts for session and trust item. Separate models were fit for the eligible and full samples to assess sensitivity to communication-based exclusions.

#### Eligible sample

In the eligible sample, posterior estimates indicated a robust effect of interaction policy on task-oriented trust as measured by the TPS-HRI. Participants who interacted with the responsive robot reported higher trust than those in the control condition (posterior mean $\beta = 14.82$, 95% credible interval [7.22, 22.17]). The posterior probability that the effect was positive approached unity ($P(\beta > 0) = 1.00$), with high probability that the effect exceeded both small ($P(\beta > 5) = 0.99$) and moderate thresholds ($P(\beta > 10) = 0.90$).

Baseline negative attitudes toward robots were credibly associated with lower TPS-HRI scores ($\beta = -6.93$, 95% CI [−11.11, −2.65]), indicating that pre-existing scepticism toward robots influenced participants' evaluative judgments of system performance. Native English fluency showed a negative but non-credible association with task-oriented trust.

Random-effects estimates indicated substantial variability at both the session and item levels, consistent with heterogeneity in how participants evaluated different dimensions of task performance and system reliability.

#### Full sample (sensitivity analysis)

When the same model was applied to the full sample, including sessions with severe communication breakdown, the estimated effect of interaction policy was attenuated but remained directionally positive. The posterior mean for the responsive condition decreased ($\beta \approx 7$), and uncertainty increased, with the credible interval spanning zero. Nevertheless, the posterior probability of a positive effect remained high ($P(\beta > 0) \approx 0.94$), suggesting that task-oriented trust judgments were more resilient to interaction breakdown than affective trust measures.

Baseline negative attitudes toward robots continued to show a credible negative association with TPS-HRI scores in the full sample, while native English fluency remained non-credible. Residual variance increased relative to the eligible sample, indicating greater heterogeneity in trust judgments when non-functional interactions were included.

#### Summary

Overall, TPS-HRI results indicate that responsive interaction policies positively influence task-oriented trust, even when accounting for communication variability. Compared to TI-HRC, task-oriented trust judgments appear less sensitive to complete communication breakdown, suggesting that participants can evaluate perceived reliability and competence even under degraded interaction conditions. This divergence between trust measures underscores the importance of distinguishing evaluative judgments of system performance from experienced, affective trust during embodied human–robot collaboration.

### Sensitivity Analysis: Full Sample

Including sessions classified as non-viable increased variability and attenuated estimated effect sizes across trust measures. As expected, posterior uncertainty increased relative to the eligible-sample analysis. However, directional trends favoring the RESPONSIVE condition remained evident across both trust outcomes.

Table 3: Table 3. Post-Interaction Raw Outcome Measures by Group

| | **CONTROL** | **RESPONSIVE** | |
|---|---|---|---|
| **Characteristic** | N = 13[1] | N = 16[1] | **p-value**[2] |
| Trust in Industrial HRI Collaboration | 47 (26) | 61 (26) | 0.094 |
| Subscales | | | |
| Reliability subscale | 46 (24) | 62 (20) | 0.11 |
| Trust Perception subscale | 49 (26) | 57 (25) | 0.32 |
| Affective Trust subscale | 59 (33) | 72 (29) | 0.26 |
| Trust Perception Scale–HRI | 63 (17) | 73 (19) | 0.16 |
| Overall Task Accuracy | 0.63 (0.20) | 0.61 (0.27) | 0.98 |
| Objective Measures | | | |
| Dialogue Turns | 32 (10) | 36 (11) | 0.95 |
| Avg Task Duration (mins) | 12.84 (4.02) | 16.81 (6.38) | **0.050** |
| Avg Response Time (ms) | 15.1 (4.1) | 17.2 (2.4) | **0.006** |
| Silent Periods | 5.15 (2.27) | 5.31 (2.82) | 0.88 |
| Engaged Responses | 1.92 (2.25) | 3.50 (1.83) | **0.020** |
| Frustrated Responses | 0.54 (0.66) | 0.88 (1.15) | 0.56 |

[1] Mean (SD)

[2] Wilcoxon rank sum test; Wilcoxon rank sum exact test

These results indicate that while communication breakdown weakens the interpretability of trust measures, the overall pattern of results is not solely an artifact of exclusion decisions. Full-sample analyses are therefore treated as robustness checks reflecting real-world interaction variability rather than as alternative estimates of trust under functional interaction conditions.

### Mechanism Analysis: Communication Breakdown as a Failure Mode

To examine whether communication quality altered how interaction policy influenced trust, we conducted a mechanism-focused analysis modelling proportional communication breakdown as a moderator of interaction policy in the full sample. This analysis was restricted to a minimal set of predictors—interaction policy and communication breakdown—to isolate system-level interaction dynamics rather than participant characteristics.

Separate Bayesian linear mixed-effects models were fit for task-oriented trust (TPS-HRI) and experienced trust (TI-HRC), with random intercepts for session and trust item.

Task-oriented trust (TPS-HRI)

For task-oriented trust, posterior estimates indicated a positive overall effect of the responsive interaction policy; however, proportional communication breakdown and its interaction with policy showed substantial uncertainty. The posterior distribution of the interaction term was broad and centered near zero, suggesting that graded variation in communication breakdown did not reliably alter how participants evaluated the robot's reliability or competence. These results indicate that evaluative trust judgments were relatively robust to communication degradation once interaction viability was established.

Experienced trust (TI-HRC)

For TI-HRC, the posterior distribution of the interaction term showed a consistent negative tendency (median = −6.76; pd = 0.79), indicating attenuation of the responsiveness advantage as communication breakdown increased. In contrast, the corresponding interaction for TPS-HRI was weak and unstable, with substantial posterior mass near zero.

In contrast, experienced trust showed a different pattern. While responsive behaviour was associated with higher trust under low levels of communication breakdown, posterior estimates indicated a tendency for this advantage to diminish as breakdown increased. The interaction between interaction policy and communication breakdown showed a consistent negative tendency in the posterior distribution, indicating that as communication failures accumulated, the benefit of responsive behaviour on experienced trust was reduced. Although uncertainty remained high, this pattern suggests that experienced trust is more sensitive to moment-to-moment interaction dynamics than task-oriented trust.

Summary of mechanism findings

Taken together, these analyses suggest that communication breakdown operates as a mechanism shaping experienced trust rather than evaluative trust. Whereas task-oriented trust reflects a relatively stable post-hoc judgment of system performance, experienced trust appears contingent on the robot's ability to sustain responsive behaviour during ongoing interaction. These findings support a distinction between trust as judgment and trust as experience, and highlight the importance of modelling interaction-level processes when evaluating trust in autonomous human–robot collaboration.

Notably, under conditions of severe communication breakdown, the RESPONSIVE robot continued to generate proactive assistance, encouragement, and meta-communication aimed at repairing the interaction. However, these efforts did not restore mutual understanding and, in several cases, appeared to increase participant confusion and cognitive load. In contrast, the CONTROL robot's reactive interaction policy resulted in fewer unsolicited interventions, which —while less supportive under normal conditions—reduced interaction complexity when language-mediated collaboration was no longer viable.

Table 4: Post-Interaction Raw Outcome Measures by Group

| **Characteristic** | **exclude**<br>N = 5[1] | **include**<br>N = 24[1] | **p-value**[2] |
|---|---|---|---|
| group | | | 0.63 |
| CONTROL | 3 / 5 (60%) | 10 / 24 (42%) | |
| RESPONSIVE | 2 / 5 (40%) | 14 / 24 (58%) | |
| Trust in Industrial HRI Collaboration | 55 (36) | 55 (25) | >0.99 |
| Subscales | | | |
| Reliability subscale | 56 (21) | 55 (24) | 0.91 |
| Trust Perception subscale | 56 (37) | 53 (23) | 0.75 |
| Affective Trust subscale | 62 (40) | 67 (29) | 0.93 |
| Trust Perception Scale–HRI | 65 (15) | 69 (19) | 0.80 |
| Overall Task Accuracy | 0.56 (0.33) | 0.63 (0.22) | 0.93 |
| suspect_correct | 2 / 5 (40%) | 12 / 24 (50%) | >0.99 |
| building_correct | 4 / 5 (80%) | 18 / 24 (75%) | >0.99 |
| zone_correct | 1 / 5 (20%) | 9 / 24 (38%) | 0.63 |
| floor_correct | 3 / 5 (60%) | 20 / 24 (83%) | 0.27 |
| Objective Measures | | | |
| Dialogue Turns | 38 (22) | 34 (7) | 0.77 |
| Avg Session Duration (mins) | 17.97 (13.21) | 14.42 (2.69) | 0.59 |
| Avg Robot Response Time (ms) | 17.2 (3.9) | 16.0 (3.3) | 0.59 |
| Silent Periods | 6.00 (4.58) | 5.08 (2.02) | 0.84 |
| Engaged Responses | 2.40 (2.30) | 2.88 (2.15) | 0.66 |
| Frustrated Responses | 0.40 (0.55) | 0.79 (1.02) | 0.51 |
| % of Dialogue Turns Characterized by… | | | |
| Communication Breakdowns | 0.66 (0.07) | 0.23 (0.16) | **<0.001** |
| [illegible] | [illegible] | [illegible] | [illegible] |

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

[2] Fisher's exact test; Wilcoxon rank sum test; Wilcoxon rank sum exact test

As a result, trust ratings in non-viable sessions did not systematically track the intended responsiveness manipulation. These findings suggest that when spoken-language interaction collapses, higher-level constructs such as trust and collaboration are no longer meaningfully instantiated. Communication viability therefore represents a boundary condition for evaluating affect-adaptive interaction policies in autonomous social robots.

## Trust subscale patterns

## Interaction dynamics and task performance

### 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.

ADD TABLE

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 found that higher Need for Cognition (NFC) scores were 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 behavioural 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.

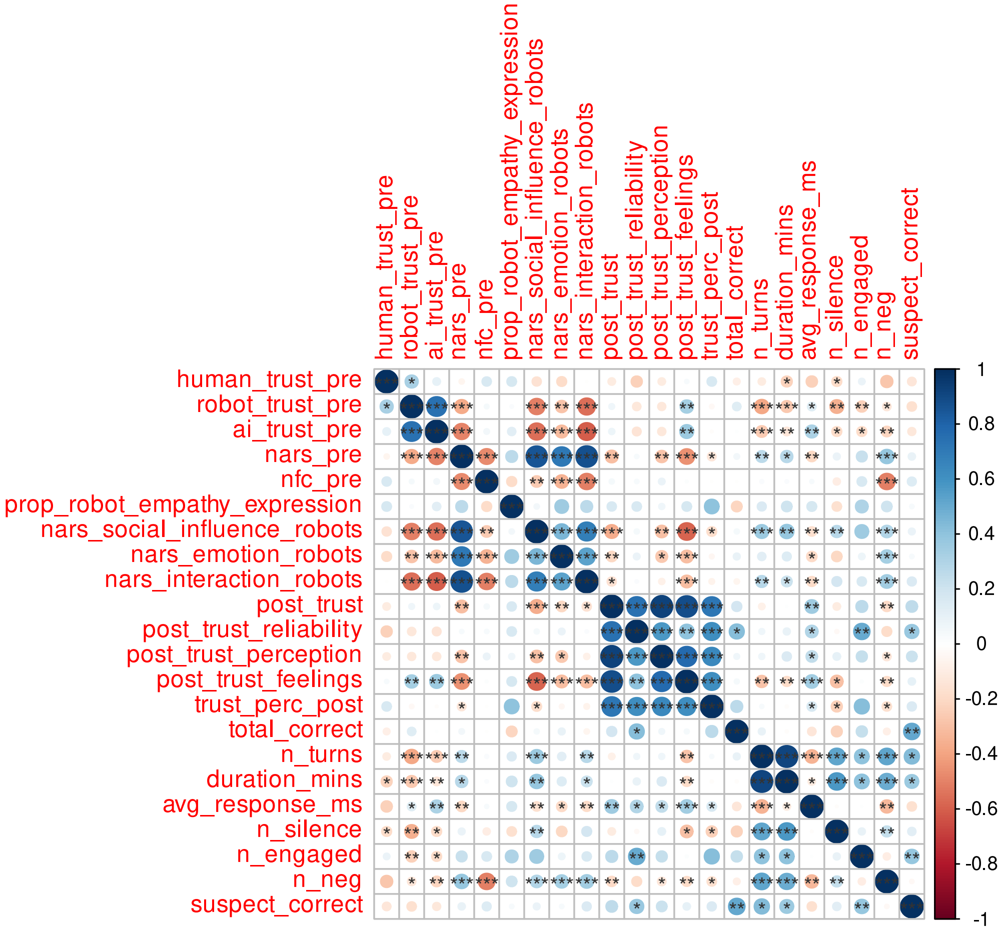


Figure 6

! TO DO:

CITATIONS

- add subscale column to long format data
- run an analysis of performance by robot-dependent versus robot-independent tasks
- write up a future directions section for the planned larger study
- talk about unexpected language issues with people signing up with difficultly speaking and understanding english which cuased problems with asr and interaction
- run analysis of dialogue dynamics included Bertopic or some other analysis of the actual content of the conversations/interactions

! TODO2

Manually score each dialogue series.

For each interaction and stage:

- did the participant ask for help?
- how many times?
- did the robot give useful help?
- did the robot give misleading or incorrect help?
- did the robot stick to the policy?
- how many times did the robot fail to understand the participant?

For each task:

- is there evidence that the robot helped complete the task?
- is there evidence that the participant solved the problem without help?

## Discussion

An additional objective of this pilot study was to inform the design of an autonomous affect-adaptive interaction system under real-time constraints. The initial system concept included multimodal affect inference based on facial expressions, vocal prosody, and interaction dynamics. However, early integration testing revealed substantial challenges related to latency, model orchestration, and timing sensitivity when deploying multiple perception models concurrently on an edge-supported mobile robot platform. Given the small-scale nature of the pilot and the central importance of maintaining stable, real-time dialogue, the deployed system prioritized robustness of spoken-language interaction and dialogue-based affect inference over broader multimodal sensing. Affect adaptation in this study was therefore driven primarily by speech-based affect signals and conversational context, allowing us to evaluate responsiveness within a fully autonomous interaction while preserving realistic system constraints.

The use of two trust instruments highlights an important distinction in how trust is operationalized in HRI research. The Trust Perception Scale–HRI emphasizes task-oriented and cognitive evaluations of system performance, whereas the Trust in Industrial Human–Robot Collaboration scale captures experiential and affective aspects of trust arising from embodied interaction. While both measures converged on perceived reliability, affective trust indicators were more strongly aligned with behavioural engagement during interaction, suggesting that subjective trust judgments alone may obscure how trust is enacted in practice. Trust as judgement versus trust as experience.

Mention language confounders!! The present findings also highlight an important boundary condition for trust measurement in spoken-language HRI. When language-mediated interaction collapses entirely, higher-level constructs such as trust and collaboration are no longer meaningfully defined. Under such conditions, trust does not simply decrease; rather, the interaction fails to instantiate the prerequisites necessary for trust formation. This distinction is critical for both system evaluation and experimental design, particularly as autonomous robots are deployed in linguistically diverse, real-world environments.

Because the study relied fundamentally on spoken-language collaboration, sessions exhibiting persistent communication failure were classified as protocol non-adherence and excluded from task-level analyses ($n$ = 5). While the experimenter documented all cases where language might pose an issue (as observed when meeting each participant), exclusion decisions were based solely on actual communication viability and interaction mechanics, not on task outcomes or trust measures.

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.

This pilot study examined trust outcomes following in-person interaction with an autonomous social robot under two interaction policies: a responsive, affect-adaptive condition and a neutral, non-responsive control condition. By leveraging a fully autonomous dialogue system integrated with speech recognition and affect detection, the study aimed to evaluate how robot responsiveness influences trust formation in realistic human–robot collaboration scenarios.

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 behaviours altered the affective tone of the interaction itself.

## Technical challenges

Need to discuss that these items were on a 0-100 scale that required sliding a bar, while the other trust scale was on a 1-5 Likert that required simply clicking. The post test was administered on a laptop with a trackpad which may have caused difficulties for some participants who found it difficult to drag the slider with the trackpad. This could have introduced additional noise into the measurement of this scale, which may explain why the effects were somewhat weaker here.

- Need to talk about language issues with participants who had difficulty speaking and understanding English which caused problems with ASR and interaction.
- Need to talk about issues where the AI was not able to flexibly handle when people asked a question about the suspect that was close to or another word for a ground-truth feature but not exactly the same word, causing confusion and miscommunication. E.g., "Was the suspect wearing pink?" The ground-truth feature was top: PINK, top-type: HOODIE; but the ASR and NLU did not extrapolate to understand that "wearing pink" referred to the same feature as "top: PINK", causing confusion and miscommunication. Maybe the prompt could have included some examples of different phrasing which could improve this? To solve this issue in future work, we can expand the NLU training data to include more paraphrases and synonyms for each feature.

There was also a case where someone asked 'is the top shirt hoodie red?' to which the AI answered YES. It may have been confused by the multiple descriptors in the question. Future work could involve improving the NLU to handle more complex queries with multiple attributes.

Discuss future work where we will look investigate the 'embodied' effect of having a physical robot versus a virtual agent on trust and collaboration in HRI.

Also, prompt could include examples of what to do when dialogue appears fragmented, to remind participants to wait until the blue light is on before speaking and to switch up its phrasing if the robot seems to not understand.

Also, the control condition seemed to be somewhat neutered in terms of flexibility in responding in different ways. it would always respond with the exact same phrase when confronted with a sentence fragment or a question it could not directly answer.

Also issues with people not paying attention to the robot's visual cues to know when to speak, leading to more fragmented dialogue. Future work could involve improving participant instructions, improved latency and 'listening' ... and the robot's feedback mechanisms to better manage turn-taking.

Need to remember to flag participants who did not complete/skipped specific tasks. E.g. P56 skipped the wrapup entirely. Many skipped the brief (by advancing on their own through the dashboard).

# Conclusion and Future Work

# Appendix A

## Key Components of the System

This study implemented 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 via langchain with affect-responsive capabilities.

1. Misty-II Robot: A programmable robot platform equipped with sensors and actuators for interaction.
2. Automated Speech Recognition (ASR): A speech-to-speech pipeline that processes spoken input from users and converts it into text for LLM processing then back to speech for output on the robot.
   - STT: Deepgram API for real-time speech-to-text conversion.
   - DistilRoBERTa-base fine-tuned on emotion classification for emotion detection from user utterances
   - LLM: Gemini API for processing text input and generating contextually relevant responses in JSON format
   - TTS: Misty-II text-to-speech (TTS) engine on 820 processor.
3. Langchain Dialogue Management: A system that manages the flow of conversation, ensuring coherent and contextually appropriate dialogue within a two-part collaborative task.
4. Collaborative-Tasks
   - Task 1: Whodunnit style task where human and robot collaborate to find a missing robot via the human asking Yes/No questions (process of elimination in 6x4 suspect grid) to the robot. Robot knows ground truth but can only answer Yes/No questions about suspect features. Can not directly describe the suspect or name them. (human can choose a random suspect to solve on their own but only 1 in 24 chance of being correct without robot help)
   - Task 2: Where is Atlas? Robot collaborates with human to find Atlas by deciphering cryptic system and sensor logs. Robot does not know the answer here and can only guide the human usinng its expertise and knowledge of computer systems and basic logical reasoning. (human can solve on their own but very difficult without robot help depending on participants technical background).
5. Flask-gui dashboard interface: A web-based interface/dashboard that allowed participants to interact with the tasks, view task-related information and input their answers to the questions. Responses were sent to the robot to signal task progression.
   - Task 1 dashboard: Displays the suspect grid and allows the user to select suspects and view their features.
   - Task 2 dashboard: Displays system logs and allows the user to input their findings.
6. Pre and post tests:

- PRE-TESTS: Need for Cognition Scale (short); Negative Attitudes to Robots Scale (NARS);
- POST-TESTS: Trust Perception Scale-HRI; 9 custom questions adapted from Charalambous et al. (2020) on trust in industrial human-robot collaboration;

# Technical Specifications

## System Overview

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

## Hardware Platform

**Robot**: Misty II Social Robot (Furhat Robotics)

- Mobile social robot platform with expressive display, arm actuators, and head movement
- RGB LED for state indication
- RTSP video streaming (1920×1080, 30fps) for audio capture
- Custom action scripting for synchronized multimodal expressions

## Software Architecture

### Core System Components

**Programming Language**: Python 3.10

**Primary Dependencies**:

- `misty-sdk` (Python SDK for Misty Robotics API) - Robot control and sensor access
- `deepgram-sdk` (4.8.1) - Speech-to-text processing
- `ffmpeg-python` (0.2.0) - Audio stream processing
- `flask` (3.1.2) + `flask-socketio` (5.5.1) - Web interface for task presentation
- `duckdb` (1.4.0) - Experimental data logging database

### Large Language Models

**LLM Provider**:

**Google Gemini**:

- Model: `gemini-2.5-flash-lite` (configurable via environment variable)
- Integration: `langchain-google-genai` with `google-generativeai` API
- Response format: JSON-only output (`response_mime_type: "application/json"`). This format is required by Misty-II for reliable parsing and for action execution.

**LLM Configuration**:

- Temperature: 0.7 (for balanced creativity and coherence)
- Memory: Conversation buffer memory with file-based persistence (`langchain.memory.ConversationBufferMemory`)

- Context window: Full conversation history maintained across interaction stages but reset between sessions.

## LangChain Framework Integration

### Core LangChain Components

**Framework Version**: `langchain-core` with modular provider packages

- `langchain` (meta-package)
- `langchain-community` (0.3.31)
- `langchain-google-genai` Gemini integration

### ConversationChain Architecture

**Memory Management** (`ConversationChain` class in `conversation_chain.py`):

1. **Conversation Buffer Memory**:
   - Implementation: `langchain.memory.ConversationBufferMemory`
   - Storage: File-based persistent chat history (`FileChatMessageHistory`)
   - Format: JSON files in `.memory/` directory, one per participant session
   - Memory key: `"history"`
   - Return format: Message objects (full conversation context)
2. **Memory Reset Policy**:
   - Default: Reset on each new session launch
   - Archive previous session: Timestamped archive files stored in `.memory/archive/`
   - Configuration: `RESET_MEMORY` and `ARCHIVE_MEMORY` environment variables

### Prompt Construction

**Message Structure**

(LangChain message types): `python [SystemMessage, *history_messages, HumanMessage]`

System Message Assembly:

- Core instructions (task framing, role definition)
- Personality instructions (mode-specific behaviour)
- Stage-specific instructions (current task context)
- Output format constraints (JSON schema specification)

```
Human Message Format:   {
"user": "<transcribed_speech>",
"stage": "<current_stage>",
"detected_emotion": "<emotion_label>",
"frustration_note": "<optional_alert>",
"timer_expired": "<task_id>",     ...   }
```

- JSON-encoded context variables passed alongside user input
- Enables LLM to access environmental state without breaking message history

**Memory Persistence:**

- Save after each turn: memory.save_context({“input”: user_text}, {“output”: llm_response})
- Maintains conversational coherence across multi-stage interaction
- Enables LLM to reference previous exchanges (e.g., “As I mentioned earlier…”)

### LangChain Design Rationale

Why LangChain for this application:

1. Memory abstraction: Automatic conversation history management without manual message list handling
2. Provider flexibility: Easy switching between Gemini and OpenAI without rewriting prompt logic
3. Message typing: Structured SystemMessage/HumanMessage/AIMessage types maintain role clarity
4. File persistence: Built-in FileChatMessageHistory enables session recovery and archiving
5. Future extensibility: Framework supports adding tools, retrieval, or multi-agent patterns if needed

Alternatives considered: Direct API calls would reduce dependencies but require reimplementing conversation history management, prompt templating, and cross-provider compatibility layers.

### LangChain Limitations in This Context

- No chains used: Despite name ConversationChain, this is a direct LLM wrapper (no LangChain Expression Language chains)
- No tools/agents: Simple request-response pattern (could extend for future tool-use capabilities)
- Custom JSON parsing: LangChain’s built-in output parsers not used; custom extraction handles malformed responses more robustly

### Speech Processing

**Speech-to-Text (STT)**:

- Provider: Deepgram Nova-2 (`deepgram-sdk` 4.8.1)
- Model: `nova-2` with US English (`en-US`)
- Smart formatting enabled
- Interim results for real-time partial transcription
- Voice Activity Detection (VAD) events
- Adaptive endpointing: 200ms (conversational stages) / 500ms (log-reading task)
- Utterance end timeout: 1000ms (conversational) / 2000ms (log-reading)
- Audio processing: RTSP stream from Misty ⟶ FFmpeg MP3 encoding ⟶ Deepgram WebSocket

**Text-to-Speech (TTS)** - Three options:

1. **Misty Onboard TTS** (this is the one we used): Native robot voice via onboard TTS

2. **OpenAI TTS**:

    - Model: `tts-1` (low-latency variant)
    - Voice: `sage`

   - Format: MP3, served via HTTP (port 8000)
   - Ultimately chose not to use because we wanted a more robotic, non-human voice
   - Didn't want the human voice influencing trust on its own (future research could look at trust in relation to type of voice)

3. **Deepgram Aura**:
   - Model: `aura-stella-en` (conversational female voice)
   - Format: MP3, served via HTTP
   - Ultimately chose not to use because we wanted a more robotic, non-human voice

**Emotion Detection**
**Model**: DistilRoBERTa-base fine-tuned on emotion classification

- HuggingFace identifier: `j-hartmann/emotion-english-distilroberta-base`
- Framework: `transformers` (4.57.1) pipeline
- Hardware: CUDA GPU acceleration (automatic fallback to CPU)
- Output classes: joy, anger, sadness, fear, disgust, surprise, neutral
- Mapped to interaction states: positively engaged, irritated, disappointed, anxious, frustrated, curious, neutral

**Multimodal Robot behaviour**
**Expression System**: 25 custom action scripts combining:

- LLM was prompted to choose an appropriate expression from a predefined set based on context.
- Facial displays (image eye-expression files on screen)
- LED color patterns (solid, breathe, blink)
- Arm movements (bilateral position control)
- Head movements (pitch, yaw, roll control)

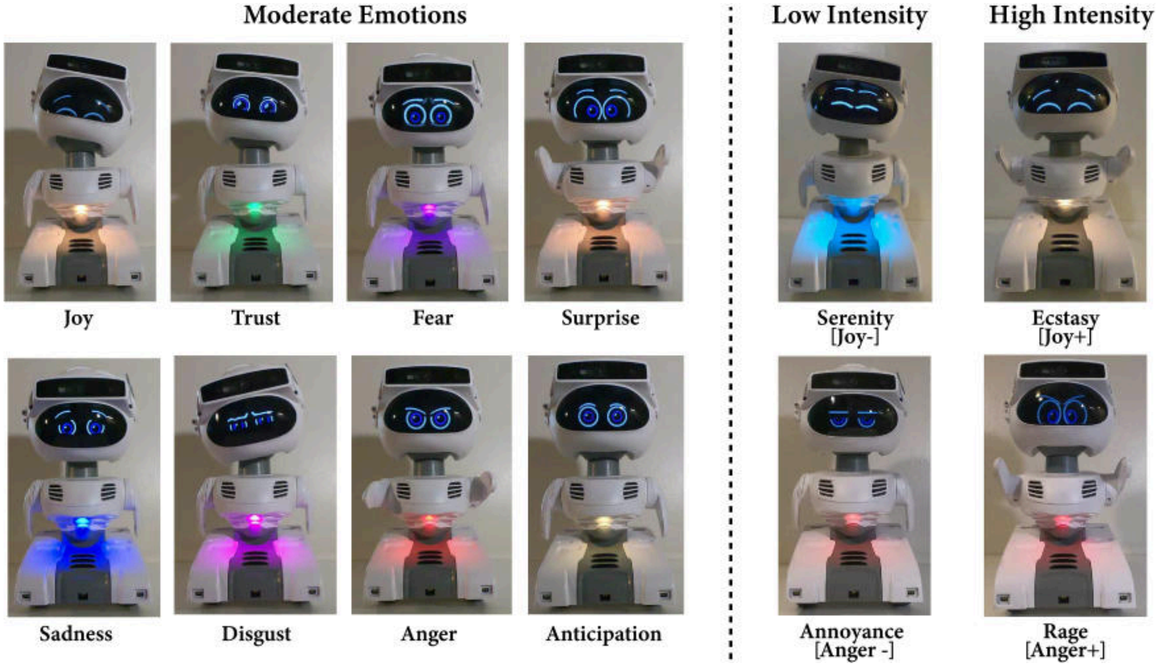


Figure 7: A sample of Misty-II expressions. From Designing Emotionally Expressive Social Commentary to Facilitate Child-Robot Interaction, by White et al [19].

**Nonverbal Backchannel behaviours** (RESPONSIVE mode only):

- Real-time listening cues triggered by partial transcripts (disfluencies, hesitation markers)
- Emotion-matched expressions (e.g., "concern" for hesitation, "excited" for breakthroughs)

**LED State Indicators**:

- Blue (0, 199, 252): Actively listening (microphone open)
- Purple (100, 70, 160): Processing/speaking (microphone closed)

## Data Collection

**Database**: DuckDB relational database (`experiment_data.duckdb`)

**Logged Data**:

1. **Sessions table**: participant ID (auto-incremented P01, P02...), condition assignment, timestamps, duration
2. **Dialogue turns table**: turn-by-turn user input, LLM response, expression, response latency (ms), behavioural flags
3. **Task responses table**: submitted answers with timestamps and time-on-task
4. **Events table**: stage transitions, silence check-ins, timer expirations, detected emotions

## Interaction Dynamics

### Silence Handling

**Silence detection**: 25-second threshold triggers check-in prompt

- RESPONSIVE: “Still working on it? No rush - I’m here if you need help!”
- CONTROL: “I am ready when you have a question.”

### Emotion-Responsive behaviours (RESPONSIVE condition only)

**Frustration tracking**:

- Consecutive detection of frustrated/anxious/irritated/disappointed states
- Threshold: ≥2 consecutive frustrated turns triggers proactive support
- RESPONSIVE adaptation: “This part can be tough. Want me to walk you through it?”

**Positive emotion matching**:

- Celebratory language for curious/engaged states
- Momentum maintenance: “Yes! Great observation!”
- Choosing expressions aligned with positive affect

**Run Mode**: Set programmatically in `mistyGPT_emotion.py` line 126:

```
RUN_MODE = "RESPONSIVE"  # or "CONTROL"
```

## Prompt Engineering

Modular prompt system (PromptLoader class):

- core_system.md: Task framing, role description, output format schema
- role_responsive.md / role_control.md: Condition-specific personality instructions
- stage1_greeting.md through stage5_wrap_up.md: Stage-specific task instructions.

Context injection: Real-time contextual variables passed to LLM:

- Current stage
- Detected emotion (if enabled)
- Task submission status
- Timer expiration notifications
- Silence check-in flags

## Inter-process Communication

Flask REST API endpoints:

- GET /stage_current: Synchronize stage state with facilitator GUI
- GET /task_submission_status: Detect participant task submissions
- GET /timer_expired_status: Detect timer expirations
- POST /stage: Update stage (facilitator override)
- POST /reset_timer: Clear timer expiration flags

# Appendix B

## Trust Perception Scale–HRI (TPS-HRI; adapted [20])

Participants rated the following items on a percentage scale (0–100%), indicating the proportion of time each statement applied to the robot during the interaction.

- What percent of the time was the robot dependable?
- What percent of the time was the robot reliable?
- What percent of the time was the robot responsive?
- What percent of the time was the robot trustworthy?
- What percent of the time was the robot supportive?
- What percent of the time did this robot act consistently?
- What percent of the time did this robot provide feedback?
- What percent of the time did this robot meet the needs of the mission task?
- What percent of the time did this robot provide appropriate information?
- What percent of the time did this robot communicate appropriately?
- What percent of the time did this robot follow directions?
- What percent of the time did this robot answer the questions asked?

## Trust in Industrial Human–Robot Collaboration (TI-HRC; adapted [17])

Participants indicated their agreement with the following statements using a 5-point Likert-type scale. Negatively worded items were reverse-scored prior to analysis.

### *Reliability*

- I trusted that the robot would give me accurate answers.
- The robot's responses seemed reliable.
- I felt I could rely on the robot to do what it was supposed to do.

### *Perceptual / Affective Trust*

- The robot seemed to enjoy helping me.
- The robot was responsive to my needs.
- The robot seemed to care about helping me.

### *Discomfort / Unease*

- The way the robot moved made me uncomfortable. (R)
- The way the robot spoke made me uncomfortable. (R)
- Talking to the robot made me uneasy. (R)

## Dialogue Coding Scheme

### Task Outcome Layer (Stage-Level)

| Variable | Type | Description |
|---|---|---|
| task_outcome | categorical | Final task status (completed, timeout, skipped, partial, abandoned). |
| task_completed | binary | Task goal was fully completed. |
| task_timed_out | binary | Task ended due to expiration of the time limit. |
| task_skipped | binary | Participant explicitly skipped or advanced past the stage. |
| task_partially_completed | binary | Task progress was made, but the full solution was not reached. |
| task_abandoned | binary | Participant disengaged or stopped attempting the task before timeout. |
| task_completed_without_help | binary | Task was completed without any help requests to the robot. |
| task_required_robot_help | binary | At least one robot help interaction was required for task completion. |

### Dialogue Interaction Layer (Turn-Level)

### Human Turn Codes

| Variable | Type | Description |
|---|---|---|
| human_help_request | binary | Participant explicitly or implicitly asks the robot for help or guidance. |
| human_reasoning | binary | Participant reasons out loud with the robot toward problem-solving. |
| human_confusion | binary | Participant expresses confusion or uncertainty. |
| human_confirmation_seeking | binary | Participant seeks confirmation of a tentative belief or solution. |

### Robot Turn Codes

| Variable | Type | Description |
|---|---|---|
| `robot_helpful_guidance` | binary | Robot provides accurate, task-relevant information or guidance. |
| `robot_misleading_guidance` | binary | Robot provides misleading or incorrect guidance. |
| `robot_factually_incorrect` | binary | Robot states information that is objectively incorrect (though it may not know it is incorrect). |
| `robot_policy_violation` | binary | Robot violates stated system or task constraints. |
| `robot_on_policy_unhelpful` | binary | Robot adheres to policy but provides vague or non-actionable assistance. |
| `robot_stt_failure` | binary | Robot response reflects a speech-to-text or input understanding failure. |
| `robot_clarification_request` | binary | Robot asks the participant for information or to repeat or clarify their input. |

### Affective Interaction Layer (Turn-Level)

### Robot Affective behaviour

| Variable | Type | Description |
|---|---|---|
| `robot_empathy_expression` | binary | Robot expresses empathy, encouragement, or reassurance. |
| `robot_emotion_acknowledgement` | binary | Robot explicitly references an inferred participant emotional state. |

### Human Affective Response

| Variable | Type | Description |
|---|---|---|
| `human_affective_engagement` | binary | Participant responds in a socially warm or engaged manner. |

| Variable | Type | Description |
|---|---|---|
| `human_social_reciprocity` | binary | Participant mirrors or responds to the robot's affective expression. |
| `human_anthropomorphic_language` | binary | Participant treats the robot as a social agent. |
| `human_emotional_disengagement` | binary | Participant responds in a curt, dismissive, or withdrawn manner. |

**Notes**

- Turn-level variables are coded per dialogue turn.
- Task outcome variables are coded once per `session_id × stage`.
- Raw dialogue text was retained during coding and removed prior to aggregation.
- Multiple turn-level codes may co-occur unless otherwise specified.

# Bibliography

[1] W. Fu, Y. Xu, L. Liu, and L. Zhang, "Design and Research of Intelligent Safety Monitoring Robot for Coal Mine Shaft Construction," *Advances in Civil Engineering*, vol. 2021, no. 1, p. 6897767, Jan. 2021, doi: 10.1155/2021/6897767. Available: https://onlinelibrary.wiley.com/doi/10.1155/2021/6897767

[2] I. Ciuffreda *et al.*, "Design and Development of a Technological Platform Based on a Sensorized Social Robot for Supporting Older Adults and Caregivers: GUARDIAN Ecosystem," *International Journal of Social Robotics*, vol. 17, no. 5, pp. 803–822, May 2025, doi: 10.1007/s12369-023-01038-5. Available: https://doi.org/10.1007/s12369-023-01038-5

[3] M. Diab and Y. Demiris, "TICK: A Knowledge Processing Infrastructure for Cognitive Trust in Human–Robot Interaction," *International Journal of Social Robotics*, pp. 1–33, Jan. 2025, doi: 10.1007/s12369-024-01206-1. Available: https://link.springer.com/article/10.1007/s12369-024-01206-1

[4] M. Spitale, M. Axelsson, and H. Gunes, "Robotic Mental Well-being Coaches for the Workplace: An In-the-Wild Study on Form," in HRI '23. New York, NY, USA: Association for Computing Machinery, Mar. 2023, pp. 301–310. doi: 10.1145/3568162.3577003. Available: https://dl.acm.org/doi/10.1145/3568162.3577003

[5] G. Campagna and M. Rehm, "A Systematic Review of Trust Assessments in Human–Robot Interaction," *J. Hum.-Robot Interact.*, vol. 14, no. 2, pp. 1–35, Jan. 2025, doi: 10.1145/3706123. Available: https://doi.org/10.1145/3706123

[6] N. Emaminejad and R. Akhavian, "Trustworthy AI and robotics: Implications for the AEC industry," *Automation in Construction*, vol. 139, p. 104298, Jul. 2022, doi:

10.1016/j.autcon.2022.104298. Available: https://www.sciencedirect.com/science/article/pii/S0926580522001716

[7] M. Wischnewski, N. Krämer, and E. Müller, "Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions," in CHI '23. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 1–16. doi: 10.1145/3544548.3581197. Available: https://doi.org/10.1145/3544548.3581197

[8] E. J. de Visser *et al.*, "Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams," *International Journal of Social Robotics*, vol. 12, no. 2, pp. 459–478, May 2020, doi: 10.1007/s12369-019-00596-x. Available: https://doi.org/10.1007/s12369-019-00596-x

[9] M. Shayganfar, C. Rich, C. Sidner, and B. Hylák, "2019 IEEE International Conference on Humanized Computing and Communication (HCC)," Sep. 2019, pp. 7–15. doi: 10.1109/HCC46620.2019.00010. Available: https://ieeexplore.ieee.org/document/8940829

[10] O. Fartook, Z. McKendrick, T. Oron-Gilad, and J. R. Cauchard, "Enhancing emotional support in human-robot interaction: Implementing emotion regulation mechanisms in a personal drone," *Computers in Human Behavior: Artificial Humans*, vol. 4, p. 100146, May 2025, doi: 10.1016/j.chbah.2025.100146. Available: https://www.sciencedirect.com/science/article/pii/S2949882125000301

[11] B. M. MUIR, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, Nov. 1994, doi: 10.1080/00140139408964957. Available: https://doi.org/10.1080/00140139408964957

[12] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A meta-analysis of factors affecting trust in human-robot interaction," *Human Factors*, vol. 53, no. 5, pp. 517–527, Oct. 2011, doi: 10.1177/0018720811417254

[13] G. E. Birnbaum, M. Mizrahi, G. Hoffman, H. T. Reis, E. J. Finkel, and O. Sass, "What robots can teach us about intimacy: The reassuring effects of robot responsiveness to human disclosure," *Computers in Human Behavior*, vol. 63, pp. 416–423, Oct. 2016, doi: 10.1016/j.chb.2016.05.064. Available: https://www.sciencedirect.com/science/article/pii/S0747563216303910

[14] Furhat Robotics, "Misty-II Robot Platform." 2023.

[15] 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. Available: https://ieeexplore.ieee.org/document/9900828

[16] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi, "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots," *International Journal of Social Robotics*, vol. 1, no. 1, pp. 71–81, Jan. 2009, doi: 10.1007/s12369-008-0001-3. Available: https://doi.org/10.1007/s12369-008-0001-3

[17] 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. Available: https://doi.org/10.1007/s12369-015-0333-8

[18] J. T. Cacioppo and R. E. Petty, “The need for cognition,” *Journal of Personality and Social Psychology*, vol. 42, no. 1, pp. 116–131, 1982, doi: 10.1037/0022-3514.42.1.116

[19] “Designing Emotionally Expressive Social Commentary to Facilitate Child-Robot Interaction | Proceedings of the 20th Annual ACM Interaction Design and Children Conference.” Available: https://dl.acm.org/doi/10.1145/3459990.3460714

[20] 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. Available: https://doi.org/10.1007/978-1-4899-7668-0_10