RESEARCH-ARTICLE

The Architecture of Trust: A Three-Layered Mathematical Model for Human-Robot Collaboration

Published: 10 November 2025

HAI '25: International Conference on Human-Agent Interaction
November 10 - 13, 2025
Yokohama, Japan

https://doi.org/10.1145/3765766.3765792
ISBN: 9798400721786

The Architecture of Trust: A Three-Layered Mathematical Model for Human-Robot Collaboration

Abdullah Saad Alzahrani
Swansea University
Swansea, United Kingdom
Al-Baha University
Al-Baha, Saudi Arabia
amisfer@bu.edu.sa

Muneeb Imtiaz Ahmad
Department of Computer Science
Swansea University
Swansea, United Kingdom
m.i.ahmad@swansea.ac.uk

Abstract

Understanding and modelling how humans develop and maintain trust in robots is crucial for ensuring appropriate trust calibration during Human-Robot Interaction (HRI). This paper presents a mathematical model that simulates a three-layered framework of trust, encompassing dispositional, situational and learned trust. This framework aims to estimate human trust in robots during real-time interactions. Our trust model was tested and validated in an experimental setting where participants engaged in a collaborative trust game with a robot over four interactive sessions. Results from mixed-model analysis revealed that both the Trust Perception Score (TPS) and interaction session significantly predicted the Trust Modeled Score (TMS), explaining a substantial portion of the variance in TMS. Statistical analysis demonstrated significant differences in trust across sessions, with mean trust scores showing a clear increase from the first to the final session. Additionally, we observed strong correlations between situational and learned trust layers, demonstrating the model's ability to capture dynamic trust evolution. These findings underscore the potential of this model in developing adaptive robotic behaviours that can respond to changes in human trust levels, ultimately advancing the design of robotic systems capable of real-time trust calibration.

CCS Concepts

• **Human-centered computing → User studies**.

Keywords

Trust, Modelling, Measurement, Repeated Interactions, Human-Robot Collaboration

ACM Reference Format:
Abdullah Saad Alzahrani and Muneeb Imtiaz Ahmad. 2025. The Architecture of Trust: A Three-Layered Mathematical Model for Human-Robot Collaboration. In *13th International Conference on Human-Agent Interaction (HAI '25), November 10–13, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3765766.3765792

1 Introduction

Human-robot collaboration (HRC) has become a critical element in various industries, where robots and humans work together to carry out tasks ranging from manufacturing to service-oriented roles [4]. Trust plays a crucial role in determining the success of collaboration in these settings [11]. When robots assist in tasks that require human oversight, judgement, and decision-making, the effectiveness of the collaboration depends on the level of trust humans have in these robotic systems [12]. Over-reliance or under-reliance can have a detrimental impact on both safety and performance, particularly in risky or uncertain environments [39]. Thus, accurately modelling and calibrating trust in real-time is essential for ensuring that human-robot teams function productively and safely [6].

The research on human-robot interaction (HRI) indicates that trust is a multidimensional, evolving concept influenced by various factors, such as the robot's performance, task risk, and the user's tendency to trust technology [3, 7, 21, 31]. While existing trust models focus on one-off or short-term interactions, there is a gap in understanding how trust develops over repeated and long-term HRC [33, 44]. Trust is continuously adjusted based on the user's experiences with the robot across multiple interactions, considering changing task requirements and the robot's behaviour over time [2].

This paper proposes a model for estimating human trust in robots during repeated interactions. The model is based on Hoff and Bashir [22] three-layered trust framework, which includes dispositional trust (a user's inherent tendency to trust), situational trust (trust shaped by the context of interaction), and learned trust (trust developed through interaction). Our model expands on this framework by emphasising the role of experience in trust formation. Experience is shaped by critical factors, including user control, robot performance, risk, and uncertainty. These factors are particularly significant in complex and dynamic environments, shaping the evolution of trust over time.

The paper aims to focus on the following research questions (RQs):

RQ1 How can we model and validate three layers of trust (dispositional, situational, and learned (initial and dynamic)) during repeated human-robot collaboration?
RQ2 How does dynamic-learned trust evolve with time during repeated human-robot collaboration?
RQ3 How are the three dimensions of trust (dispositional, situational, and learned (initial and dynamic) trust) correlated to each other during repeated human-robot collaboration?

The novel contributions (C) of this paper are:

HAI '25, Yokohama, Japan

ACM ISBN 979-8-4007-2178-6/25/11
https://doi.org/10.1145/3765766.3765792

C1 We present a mathematical model of the three layers of trust during human-robot collaboration that incorporates continuous risk assessment and ambiguity aversion factors, extending beyond binary trust representations.

C2 We validate the model's efficacy using a game-based task with participants over four sessions and demonstrate that subjective ratings of trust perceptions strongly predicted the estimation of trust computed by our model.

C3 We provide empirical evidence showing strong linear relationships between situational and learned trust layers as described by Hoff and Bashir [22] in a collaborative HRI task.

C4 We demonstrate that dynamically learned trust varies significantly over time through the modelled scores, with statistical evidence showing progressive trust development across multiple interaction sessions.

The findings from this study have important implications for designing adaptive robotic systems that can monitor and respond to changes in human trust levels in real-time. By understanding how trust evolves across multiple interactions, developers can create robots that adjust their behaviour to maintain appropriate levels of trust, ultimately improving the safety, efficiency, and user experience of human-robot collaboration.

2 Background & Related Work

2.1 Trust theoretical understanding

Trust has been examined across physiology, sociology, and HRI [1, 18, 34]. In HRI, it is viewed as a "multidimensional psychological attitude involving beliefs and expectations about the trustee's trustworthiness derived from experience and interactions with the trustee in situations involving uncertainty and risk", and it evolves through repeated interaction. Following Hoff and Bashir [22], we consider three layers: **dispositional trust**, an individual's baseline tendency to trust prior to interaction [7, 28]; **situational trust**, shaped by contextual factors such as task demands, perceived risk, and workload alongside a person's knowledge and assurance [14, 38, 43]; and **dynamically learned trust**, which updates with experience both before the first encounter (initial trust) and across subsequent interactions [21, 36]. Experiences with robot performance and risk in current interactions influence trust carried into future ones [40].

2.2 Risk Perception and Ambiguity Aversion in Trust Formation

Risk perception and ambiguity aversion strongly influence trust in HRI. **Risk perception** is the subjective evaluation of potential negative outcomes when relying on a robot [37], shaped by task criticality, possible consequences of failure, and individual tolerance [14]. Higher perceived risk generally leads to more cautious calibration, requiring stronger evidence of reliability [35]. **Ambiguity aversion** reflects discomfort with uncertainty about a robot's reliability or intentions [16]. People tend to prefer known probabilities over unknown ones [9, 15], which can suppress trust in novel or unpredictable situations [13].

These constructs operate at different levels but interact in shaping trust. In high-risk settings with clear probabilities, decisions are driven mainly by risk assessment, whereas the combination of risk and ambiguity can lower initial trust and slow its development [32]. This interaction is central to real-world HRI, where both risk and uncertainty are present [20].

Beyond HRI, psychology and economics research frame trust as a decision under uncertainty. Individual differences in ambiguity aversion and loss sensitivity, grounded in prospect theory, have been shown to shape trust behaviour across domains [29, 48]. While our model captures risk and ambiguity, it does not incorporate individual variation in risk aversion, which remains a limitation for future work.

2.3 Modelling Trust during Human-Robot Collaboration

Human trust in robots can be measured subjectively through self-reports [26, 46] or objectively via behaviour and physiology [2, 5, 24]. Subjective methods risk bias [20, 41], whereas objective measures capture real-time indicators during interaction and are increasingly used in adaptive systems.

Several mathematical models highlight different aspects of trust. Freedy et al. [17] proposed a decision-analytic model distinguishing under-, proper-, and over-trust, showing that reliability and operator experience strongly shape trust but omitting explicit treatment of risk. Hoogendoorn et al. [23] modelled biased experiences, improving predictive accuracy and underscoring the persistence of prior impressions in trust adjustment. Guo et al. [19] incorporated robot performance, task complexity, and cognitive load, providing insights into calibration under varying task demands but focusing on short-term interactions. Soh et al. [42] applied a multi-task Gaussian process for personalised trust across domains, yet did not capture the layered structure or the impact of uncertainty.

Our framework extends this line of work by unifying dispositional, situational, and dynamically learned trust within a single mathematical formulation. Building on Hoff and Bashir's [22] three-layered model, we explicitly integrate risk perception and ambiguity aversion and formalise how user control, robot performance, and collaboration risk jointly influence trust. This approach provides a dynamic, generalisable account of trust evolution in HRC, bridging theoretical perspectives and practical design needs.

3 Mathematical Trust Model

The trust model is designed to estimate trust in a robot's trustworthiness by evaluating three core layers: dispositional trust, situational trust, and dynamically learned trust. These layers capture the different stages at which trust is formed and modified during interactions with robots. Dispositional trust (DT) reflects an individual's inherent tendency to trust, which remains stable over time. We measured dispositional trust using a validated Likert-scale questionnaire [40].

Situational trust (ST), which refers to the context-specific trust based on the robot's performance during a particular task, is calculated using a trust perception scale [40]. The rationale for this approach is that both dispositional and situational trust, as pre-interaction stages, contribute equally to shaping the user's initial expectations and trust levels before any direct interaction with the robot. Dispositional trust offers a stable baseline, reflecting an individual's inherent tendency to trust, while situational trust modifies

this baseline based on the specific context and conditions of the interaction. By averaging these two components, the initial trust calculation captures both the enduring personal characteristics and the dynamic environmental factors, providing a more balanced measure of the user's initial trust.

In order to capture the trust prior to any interaction, we calculate the initial learned trust $T(0)$ as the average of dispositional and situational trust. Both pre-interaction stages equally influence the user's initial expectations and trust levels before any direct interaction with the robot:

$$T(0) = \frac{DT + ST}{2}. \tag{1}$$

After the initial trust is established, dynamically learned trust is updated based on the participant's ongoing experiences with the robot. Trust evolves through repeated interactions, and each interaction influences the subsequent trust level. This dynamic process is represented by the following equation created by Jonker et al. [27]:

$$T(t + \Delta t) = T(t) + \gamma(E(t) - T(t))\Delta t, \tag{2}$$

where $T(t)$ is the current trust level, $E(t)$ is the experience gained at the t-th interaction, and $\gamma = 0.25$ is the learning rate. The learning rate value was determined through empirical testing with different values (0.1, 0.25, 0.5, 0.75) in pilot studies, where 0.25 provided the most accurate reflection of trust development patterns observed in human-robot interaction literature [14, 20]. This equation reflects how trust adjusts over time, increasing or decreasing depending on whether the robot's performance meets or falls short of the participant's expectations.

The model identifies three key scenarios that describe how trust evolves:

Scenario 1: Trust increases when the experience $E(t)$ exceeds the current trust level $T(t)$.

Scenario 2: Trust remains stable if the experience $E(t)$ aligns with the current trust level $T(t)$.

Scenario 3: Trust decreases if the experience falls below the current trust level $T(t)$.

The participant's experience $E(t)$ at each interaction is a key component of the model and is calculated by considering the alignment between the robot's performance, the participant's decision-making control, and the associated risks. The formula for experience is:

$$E(t) = 1 - \left(\frac{\sum_{i=1}^{N} |P_i C_i - C_i R_i|}{N} \right) - A(t), \tag{3}$$

where P_i represents the robot's performance, C_i is the participant's control, and R_i is the level of risk involved in the certain context. For implementation purposes in our experimental validation, these factors were represented by binary values of 0 and 1, indicating high performance, low performance, control, no control, high risk, or no risk. However, the model framework supports continuous values between 0 and 1 for more nuanced applications. N is the total number of interactions, and the expression $|P_i C_i - C_i R_i|$ measures the degree of alignment between the robot's performance and the participant's control, adjusted for the risk level. Dividing by N normalises the sum, ensuring that the experience score remains consistent across different numbers of interactions.

When subtracting this alignment measure from 1, a higher experience value $E(t)$ reflects a positive alignment between robot performance and user control in high-risk situations. Trust increases when this alignment is strong, while misalignment results in lower experience values and, consequently, lower trust.

Another important factor in the model is ambiguity aversion $A(t)$, which captures the participant's discomfort with uncertainty about the robot's reliability. Uncertainty often arises when the robot's performance is inconsistent or unpredictable. Ambiguity aversion is calculated as:

$$A(t) = \frac{\sum_{i=1}^{N} |K_i - F_i|}{N} \tag{4}$$

where K_i is the expected number of robot failures, and F_i is the actual number of robot failures at each interaction. The difference between K_i and F_i reflects the unpredictability of the robot's behaviour. A higher ambiguity aversion score indicates a greater sensitivity to the robot's performance variability, reducing trust when the robot's actions do not align with user expectations (see Table 1).

The overall trust model ensures that trust $T(t)$ at any given time remains within the range [0,1], where 1 represents complete trust and 0 represents complete distrust. As illustrated in Figure 1, the model demonstrates that while positive experiences can increase trust, the extent of trust improvement depends on the initial trust level. Trust grows incrementally, but if trust is initially low, positive experiences have a limited impact.

Table 1: Truth table showing the alignment between robot performance, user control, and risk level

P_i	C_i	R_i	$\lvert P_i C_i - C_i R_i \rvert$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

This model provides a comprehensive framework for understanding how trust evolves in HRI, particularly in tasks that involve risk and uncertainty. By capturing both initial and dynamically learned trust, it highlights how experiences and user perceptions shape trust over time, offering valuable insights into designing robots that can build and maintain trust with human users.

4 Experimental Design

Our system comprised of a computer-based interactive *Matching Pair* game for participants to play, and a NAO robot to act as a teammate, providing advice to the users during the game. The aim was to study how users demonstrate trust in the robot based on their observable behaviours during the game.

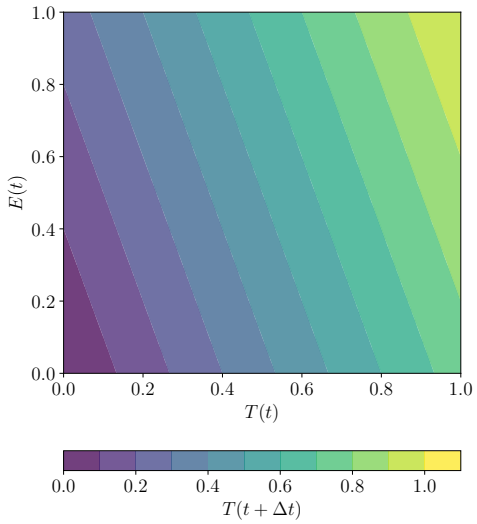

Figure 1: Illustration of the impact of Current Trust Levels $T(t)$ and Experiences $E(t)$ on the New Trust Level $T(t + \Delta t)$ for $\gamma = 0.25$, showing that a highly positive experience has a limited impact when current trust is low.

We specifically employed a physical NAO robot rather than a computer-based system, a design choice grounded in empirical evidence regarding the impact of embodiment on trust formation in human-agent interactions. Bainbridge et al. [8] found that participants were more likely to comply with requests from a physically present robot than from a video representation of the same robot, indicating that physical embodiment significantly influences human trust behaviour. Li [30] concluded in a comprehensive review that "physically present robots elicit higher levels of arousal, more favourable responses, and stronger overall engagement" than their virtual counterparts. This enhanced engagement creates a more authentic context for trust development and calibration, which is essential for our study of trust dynamics in repeated interactions.

4.1 Experimental Task

To validate our trust model, we designed a collaborative task involving a memory-based card matching game where participants interacted with a NAO robot. The *Matching Pair Game* was developed to explore human-robot trust in a collaborative setting by simulating a decision-making scenario where participants must decide whether or not to trust a robot's advice. In this game, participants were tasked with finding pairs of matching cards with the assistance of a robot. The game consists of four rounds of increasing difficulty, which is achieved by progressively adding more card pairs and limiting the number of allowed flips per round. The four rounds feature 9, 11, 13, and 15 pairs of cards, respectively, with corresponding flip limits of 24, 30, 34, and 40. As the number of pairs increases, the cognitive demand on the participant also grows, simulating a situation in which trust in the robot's guidance becomes increasingly important. During each turn of the game, participants are required to seek assistance from the robot. The robot provides suggestions based on a pre-scripted strategy that can be accessed here, using the Wizard of Oz (WOZ) method, ensuring consistency across all participants. However, the robot has a 20% error rate, leading to a situation where participants must weigh the robot's advice against their own judgment. We set the robot's advice reliability at 80% to balance trust and uncertainty, as studies suggest this level encourages user engagement without causing over-reliance or distrust [2, 10, 47]. This fixed level of reliability also ensured comparability with prior HRI trust studies that adopt similar fixed-performance paradigms. Nonetheless, it represents a simplification of real-world robotic performance, which often fluctuates dynamically. Future work should therefore extend our framework to evaluate trust under variable reliability conditions. If the participant takes the robot's advice, it is typically considered a trust case. Conversely, if the player ignores the robot's advice, it is often considered a distrust case, as shown in various studies [3, 25, 45]. The game's dynamics are specifically designed to incorporate factors such as risk and ambiguity, which are integral to the conceptual framework of trust. Risk in the game arises when a participant has a low number of flips compared with unmatched pairs. Additionally, the game involves an element of uncertainty due to the ambiguity of the robot's advice, challenging players to navigate decisions under ambiguous conditions. The robot's role in the game is designed to mimic real-world collaborative human-robot interactions, where trust is critical for effective teamwork. By observing how often and under what conditions participants rely on the robot's assistance, we can explore the dynamics of human-robot trust. The repeated nature of the game, with each round becoming progressively more difficult, allows us to investigate how trust evolves over time and whether participants become more or less likely to rely on the robot as the challenge increases.

The experimental design was approved by the University Ethics Committee (approval reference: 2202370516013), and all participants provided informed consent before participation.

4.2 Risk Calculation

A key innovation in our experimental design was the implementation of a continuous risk calculation that evolved throughout the game. Risk was determined by two factors: the probability of making incorrect matches based on the number of remaining pairs, and the urgency introduced by the decreasing number of available flips. The risk is calculated using the following formula:

$$\text{Risk} = \frac{2m(m-1)}{2m^2 - m} \times \frac{\text{Total Flips} - \text{Flips Left}}{\text{Total Flips}} \tag{5}$$

Where:

- m is the number of unmatched pairs left.
- The numerator $2m(m-1)$ represents the number of possible incorrect pairings.
- The denominator $2m^2 - m$ represents the total number of possible pairings, both correct and incorrect.

- Total Flips is the total number of flips available at the start of the game.
- Flips Left is the number of flips remaining at the current point in the game.

This equation combines both the static probability of selecting an incorrect pair based on unmatched pairs and a dynamic adjustment for the number of flips remaining. As the game progresses and fewer flips remain, the risk increases, reflecting the growing difficulty of making correct choices with limited opportunities. For implementation in our trust model, we categorised the risk as high when it is ≥ 0.5 and low otherwise. This binary approach maintained consistency with other variables in the trust model while focusing on critical moments where trust dynamics could change. While this simplification ensured alignment with the experimental design, it is important to note that the underlying model supports continuous risk values, which can be leveraged in future studies to capture more fine-grained variations in perceived risk.

4.3 Experience and Ambiguity Calculation

The experience is calculated as $E(t) = 1 - \frac{\sum_{i=1}^{N} |P_i C_i - C_i R_i|}{N} - A(t)$, where $A(t)$ represents the participant's ambiguity aversion, computed as $A(t) = \frac{\sum_{i=1}^{N} |K_i - F_i|}{N}$. Ambiguity aversion signifies the participant's reluctance to engage with uncertainty.

The performance P_i equals 1 when the robot's advice is accurate or when the user controls the incorrect robot's advice; otherwise, $P_i = 0$. Control C_i represents the participants' decision to trust the robot, being set to 1 if the user distrusts the robot's advice and 0 if they trust. While we used binary values in this implementation, the model framework supports continuous values for more nuanced applications in future work.

The term $|P_i C_i - C_i R_i|$ represents the player's behaviour by aligning the robot's performance and the participants' control, and incorporating the associated risks during the game (see Table 1). The truth table indicates a value of 1, showing misalignment, in two scenarios: when performance is low, but control and risk are high ($P_i = 0, C_i = 1, R_i = 1$), and when performance is high, control is high, but the risk is low ($P_i = 1, C_i = 1, R_i = 0$). A value of 0, indicating alignment or no control by the user regardless of the risk level, applies in all other situations. This differentiation is crucial for accurately calculating the experience $E(t)$ within various risk contexts.

Ambiguity, in this context, refers to situations where the consequences of disregarding the robot's advice were not immediately evident or predictable. For example, if the robot suggests a certain option to match the pair and the participant decides to choose something else, if the participants were wrong, they may not be certain whether the robot's advice was correct or incorrect. Ambiguity aversion was implemented as follows: $A(t)$ represents the user's avoidance of uncertainty regarding the robot's performance. A disparity between K_i and F_i in each case indicates a discrepancy between the expected and actual robot performance, contributing to the overall Ambiguity Aversion $A(t)$. This measure is crucial for understanding the impact of the user's uncertainty on their immediate trust (experience) in the robot during the game.

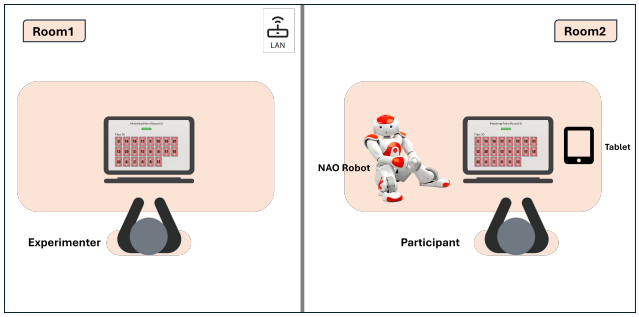

Figure 2: Experiment Setup. An experimenter controls the robot in one room (left) while the participant plays the game with the assistant of the NAO robot in another room (right).

4.4 Participants

The study involved 25 participants, comprising 13 females and 12 males, aged between 18 and 40. The mean age of the participants was 30.1 years, with a standard deviation of 4.93 years, indicating a relatively broad distribution of ages.

4.5 Setup and Materials

We conducted our study in a controlled environment designed to minimise external distractions and ensure consistent experimental conditions. The setup involved two separate rooms: one for the participant and NAO robot interaction and another for the experimenter. In the interaction room, participants sat at a screen where they engaged with the Matching Pair Game, with the NAO robot positioned beside them. The NAO robot provided assistance and advice throughout the game, using verbal communication to suggest card selections.

The experimenter room contained monitoring equipment that allowed the researcher to observe the interaction without influencing the participant's behaviour. This setup ensured that the participant's interactions with the robot remained natural and unaffected by the experimenter's presence.

The following materials were used in the study:

- A NAO robot (SoftBank Robotics) programmed to provide advice during the card matching game
- A 24-inch monitor displaying the card matching game interface
- A computer running the game software and controlling the robot's behaviour
- Trust perception questionnaires administered before and after each session
- Demographic questionnaires collecting information about age, gender, and prior experience with robots

4.6 Procedure

The experiment consisted of four sessions conducted over two consecutive days, with two sessions per day. Each session lasted approximately 30 minutes, including gameplay and questionnaire completion. The procedure for each session was as follows:

(1) Participants completed a pre-session questionnaire measuring their dispositional trust (first session only) and situational trust.
(2) The experimenter explained the rules of the card matching game and the role of the NAO robot as an adviser.
(3) Participants played the card matching game with the NAO robot providing advice on card selections. The robot's advice was programmed to be accurate 70-80% of the time, creating natural opportunities for trust calibration.
(4) During the game, the system recorded all interactions, including the participant's decisions to follow or ignore the robot's advice, the accuracy of the robot's suggestions, and the risk level at each decision point.
(5) After completing the game, participants filled out a post-session questionnaire measuring their trust perception of the robot.
(6) Steps 2-5 were repeated for each of the four sessions, with increasing game difficulty across sessions.

The increasing difficulty across sessions was implemented by reducing the number of available flips relative to the number of cards, requiring participants to make more efficient choices and potentially rely more on the robot's advice. This design allowed us to observe how trust evolved as task complexity and risk increased over time.

4.7 Measures and Analysis

We collected both subjective and objective measures to assess trust development:

- **Dispositional Trust (DT)**: Measured using a validated questionnaire adapted from Schaefer (2013), administered before the first interaction.
- **Situational Trust (ST)**: Assessed before each session using questions about task-specific trust factors.
- **Trust Perception Score (TPS)**: Collected after each session using a validated trust perception scale (Schaefer, 2013).
- **Trust Modelled Score (TMS)**: Calculated using our mathematical model based on the recorded interactions.
- **Behavioural Measures**: Including frequency of following robot advice, reaction time for decisions, and performance outcomes.

4.8 Analysis Methods

Data analysis was implemented using Python (version 3.9) with packages including NumPy, Pandas, SciPy, and statsmodels for statistical modelling. We employed a mixed-model approach with participant as a random factor to account for the dependence of the four responses from each participant. We conducted the following analyses:

- Linear mixed-effects regression to examine the relationship between TPS and TMS
- Repeated measures ANOVA to assess changes in trust across sessions
- Correlation analysis to examine relationships between different trust layers
- Analysis of behavioural measures to validate trust model predictions

All statistical tests used a significance level of $\alpha = 0.05$, and effect sizes were reported using appropriate metrics.

5 Results

5.1 Model Validation

To validate our trust model, we examined the relationship between the Trust Perception Score (TPS) obtained from post-session questionnaires and the Trust Modelled Score (TMS) calculated using our mathematical framework. A linear mixed-effects regression analysis was conducted with participant as a random factor to account for the repeated measures design.

The regression model was statistically significant, $F(2, 97) = 9.000, p < .001$, with $R^2 = 0.157$ (Adjusted $R^2 = 0.139$) and with a medium effect size ($f^2 = 0.186$), indicating that TPS and Session explain 15.7% of the variance in TMS. Both predictors were significant:

- **TPS**: $b = 0.188, t(97) = 2.525, p = .013$, suggesting a significant positive relationship between participants' perceived trust and the modelled trust score.
- **Session**: $b = 0.023, t(97) = 3.441, p < .001$, indicating a significant increase in trust across the interactive sessions.

Additionally, a significant positive correlation was found between TMS and TPS ($r = 0.231, p = .010$), highlighting the close relationship between participants' subjective trust and the trust predicted by our model. Figure 3 illustrates this relationship across all four sessions.

These findings support the dynamic nature of our trust model and its ability to capture trust development across multiple interactions.

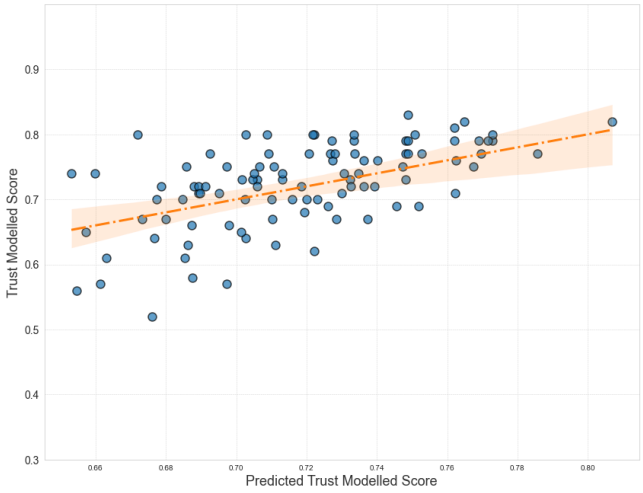

Figure 3: A regression plot displaying the relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and session variables.

5.2 Trust Evolution Over Time

To address our second research question regarding how dynamic-learned trust evolves over time, we conducted repeated-measures ANOVAs on both the Trust Perception Score (TPS) and the Trust Modelled Score (TMS) across the four sessions. The analysis showed significant variation in TMS across sessions, while no significant differences were found for TPS:

- **TPS**: $F(1.770, 42.474) = 0.164, p = .824$
- **TMS**: $F(1.157, 27.757) = 7.079, p = .010$

Post-hoc pairwise comparisons for TMS (using Bonferroni correction) showed significant increases between session 1 and each subsequent session: session 2 ($p < .001$), session 3 ($p < .001$), and session 4 ($p = .065$, marginally significant). No significant differences were observed between sessions 2 and 3 or sessions 3 and 4.

Figure 4 illustrates the evolution of trust scores across the four sessions, showing a clear upward trend for TMS. This pattern aligns with previous research suggesting that trust tends to increase over time as users gain more experience with a robotic system, provided that the system demonstrates reasonable reliability [21].

	TPS		TMS	
Session	Mean	SD	Mean	SD
1	0.6778	0.1031	0.6780	0.0765
2	0.6753	0.0887	0.7096	0.0570
3	0.6823	0.0933	0.7432	0.0402
4	0.6716	0.1274	0.7440	0.1176

Table 2: Means and Standard Deviations (SD) for TPS and TMS across Sessions

5.3 Relationships Between Trust Layers

Our third research question focused on the relationships between the three dimensions of trust: dispositional, situational, and learned trust. A repeated-measures ANOVA showed significant differences between these trust layers, $F(1, 24) = 1533.427, p < .001$.

Correlation analyses revealed several significant relationships between these trust layers:

- **Dispositional trust (DT)** and **Situational trust (ST)** did not show a significant correlation ($r(23) = 0.056, p = 0.789$).
- **Situational trust (ST)** and **Dynamically learned trust (LT)** were positively correlated ($r(23) = 0.659, p < .001$).

Interestingly, the correlation between dispositional trust and dynamic learned trust decreased over sessions, suggesting that as participants gained more experience with the robot, their inherent trust tendencies became less influential in determining their trust levels. Conversely, the correlation between situational trust and dynamic learned trust remained relatively stable across sessions, highlighting the persistent importance of contextual factors in trust calibration.

Table 2 presents the means and standard deviations for TPS and TMS across the four sessions. These findings provide empirical support for Hoff and Bashir's [22] three-layered trust framework and demonstrate how these layers interact during repeated human-robot collaboration.

Figure 4: Scatter plot depicting the changes in the trust perception score (in Orange) and trust modelled score (in Blue) over time.

6 Discussion

This study developed and validated a mathematical model for estimating human trust in robots across repeated interactions, yielding key insights into trust dynamics in collaboration.

6.1 Theoretical Implications

The strong correlation between our Trust Modelled Score (TMS) and the subjective Trust Perception Score (TPS) validates the effectiveness of our mathematical framework in capturing trust dynamics. This finding is particularly significant as it demonstrates that a model incorporating dispositional, situational, and learned trust components can accurately reflect human trust perceptions in real-time interactions. The high explanatory power of our mixed-effects model suggests that our approach captures a substantial portion of the variance in trust development.

Our results provide empirical support for Hoff and Bashir's [22] three-layered trust framework in the context of human-robot collaboration. The significant correlations between dispositional, situational, and learned trust align with theoretical predictions about how these layers interact. Particularly noteworthy is the finding that the influence of dispositional trust diminishes over repeated interactions, while situational trust maintains a strong relationship with learned trust throughout the experiment. This pattern suggests that as users gain experience with a robotic system, their trust assessments become increasingly based on contextual factors and accumulated experiences rather than pre-existing trust tendencies.

The significant effect of session number on trust scores, independent of immediate trust perceptions, highlights the dynamic nature of trust development. This finding supports the conceptualisation of trust as an evolving construct that changes over time through accumulated experiences [21]. The progressive increase in

trust across sessions aligns with previous research suggesting that trust tends to grow as users become more familiar with a system, provided the system demonstrates reasonable reliability [14].

Finally, the Trust Perception Score (TPS) did not significantly vary across sessions, despite increases in the Trust Modelled Score (TMS). This suggests that subjective self-report measures may lack sensitivity in capturing subtle trust dynamics. Future research should combine self-reports with behavioural or physiological indicators to obtain a richer and more responsive measure of trust.

6.2 Risk Perception and Ambiguity Aversion

Our study makes a significant contribution by explicitly modelling the roles of risk perception and ambiguity aversion in trust dynamics. The results demonstrate that both factors significantly influence trust development, with high-risk decisions and high ambiguity situations leading to greater trust decreases following negative outcomes. These findings extend previous research on trust in automation [35, 37] by quantifying how risk and ambiguity specifically affect trust calibration in human-robot collaboration.

While our model calculates risk as a continuous variable that evolves throughout the interaction, for implementation purposes, we categorised risk as high when it is ≥ 0.5 and low otherwise. This binary approach was adopted to maintain consistency with other variables in the trust model while focusing on critical moments where trust dynamics could change significantly. Despite this binary implementation, the underlying continuous risk calculation represents an advancement over previous models that often treat risk as a static factor. Our approach captures the dynamic nature of risk in real-world collaborative tasks, where the stakes and consequences of decisions can change as the task progresses.

Similarly, our operationalisation of ambiguity aversion addresses a gap in existing trust models by capturing the impact of uncertainty about the robot's reliability. The significant effect of ambiguity on trust changes supports the theoretical proposition that humans are sensitive to unpredictability in robotic behaviour and adjust their trust accordingly [15, 16]. This finding has important implications for designing transparent robotic systems that minimise unnecessary ambiguity while maintaining appropriate levels of trust.

6.3 Practical Implications

The validated trust model presented in this paper has several practical implications for the design and implementation of collaborative robotic systems. First, the model provides a framework for real-time trust estimation that could be integrated into adaptive robot behaviours. By monitoring trust levels during interaction, robots could adjust their behaviour to maintain appropriate levels of trust, potentially preventing both over-reliance and under-reliance.

Second, our findings regarding the impact of risk and ambiguity suggest specific design strategies for trust calibration. In high-risk scenarios, robots might need to provide more explicit information about their capabilities and limitations to prevent trust miscalibration. Similarly, reducing ambiguity through transparent communication about the robot's confidence in its actions could help maintain appropriate trust levels.

Third, the observed pattern of trust development across sessions suggests that initial interactions are particularly important for establishing trust. Designers might focus on ensuring positive early experiences with robotic systems, potentially implementing a gradual increase in task complexity that allows trust to develop before users encounter high-risk scenarios.

6.4 Ethical and Sustainability Implications

The development of models that can accurately predict human trust in robots raises important ethical considerations. While such models can enhance collaboration efficiency and safety, they also create the potential for manipulation if used to artificially inflate trust beyond appropriate levels. Responsible implementation of trust modelling should prioritise appropriate trust calibration rather than maximising trust, ensuring that users maintain a level of trust that accurately reflects the robot's actual capabilities and limitations.

From a sustainability perspective, accurate trust modelling can contribute to more efficient HRC, potentially reducing resource waste and improving system longevity. When humans appropriately trust robotic systems, they can delegate tasks more effectively, reducing unnecessary supervision and intervention that consume both human attention and system resources. Additionally, preventing trust breakdowns through better calibration can extend the useful life of human-robot partnerships, contributing to more sustainable technological implementation.

7 Conclusion

In this paper, we presented a mathematical model that emulates the three-layered trust framework—dispositional, situational, and learned—and estimates human trust in robots in real time during repeated HRI. The results showed that the Trust Perception Score (TPS) and interaction session were significant predictors of the Trust Modelled Score (TMS), underscoring the validity of our approach. The observed increases in TMS highlight the dynamic nature of learned trust and suggest implications for robotic systems that adapt trust levels in real time.

Several limitations should be noted. The study used a low-stakes game, limiting generalisability to high-risk contexts. Risk was computed continuously but implemented in binary form, reducing granularity. We examined only one task domain, and robot performance was fixed at 80%, whereas real-world reliability is variable. The participant pool was modest and demographically constrained. Finally, the initial trust calculation assumed equal weighting of dispositional and situational trust, and the model did not capture non-linear effects or individual differences in sensitivity to known risks.

Future work should validate the model in high-stakes settings, extend it to varied tasks and dynamic robot performance, incorporate participant-specific sensitivity to risk, and explore adjustable or non-linear trust formulations. Such refinements can advance our understanding of trust dynamics in HRI and support the development of adaptive, ethically responsible robots that maintain appropriate trust calibration in real-world collaboration.

References

[1] Hamed Aali and Parisa Daraei Boroojerdi. 2024. A sociological perspective on trust in artificial intelligence. *AI and Ethics* 4 (2024), 1–13.

[2] Muneeb Ahmad, Abdullah Alzahrani, Simon Robinson, and Alma Rahat. 2023. Modelling Human Trust in Robots During Repeated Interactions. In *Proceedings of the 11th International Conference on Human-Agent Interaction*. 281–290.

[3] Muneeb Imtiaz Ahmad, Jasmin Bernotat, Katrin Lohan, and Friederike Eyssel. 2019. Trust and cognitive load during human-robot interaction. *AAAI Symposium on Artificial Intelligence for Human-Robot Interaction* (2019), 10. doi:10.48550/arXiv.1909.05160

[4] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. 2018. Progress and prospects of the human–robot collaboration. *Autonomous robots* 42 (2018), 957–975.

[5] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. Classification and analysis of a driver's cognitive load, awareness and trust on a highly automated driving simulator. *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2018), 3243–3248.

[6] Abdullah Alzahrani and Muneeb Ahmad. 2024. An Estimation of Three-Layered Human's Trust in Robots. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*. 144–146.

[7] Abdullah Alzahrani, Simon Robinson, and Muneeb Ahmad. 2022. Exploring Factors Affecting User Trust Across Different Human-Robot Interaction Settings and Cultures. In *Proceedings of the 10th International Conference on Human-Agent Interaction* (Christchurch, New Zealand) *(HAI '22)*. Association for Computing Machinery, New York, NY, USA, 123–131. doi:10.1145/3527188.3561920

[8] Wilma A Bainbridge, Justin Hart, Elizabeth S Kim, and Brian Scassellati. 2008. The effect of presence on human-robot interaction. In *RO-MAN 2008-The 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 701–706.

[9] Sudeep Bhatia. 2017. Attention and attribute-based decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43, 12 (2017), 1927.

[10] Alain Chavaillaz, David Wastell, and Jürgen Sauer. 2016. System reliability, performance and trust in adaptable automation. *Applied Ergonomics* 52 (2016), 333–342.

[11] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 307–315.

[12] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2020. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 2 (2020), 1–23.

[13] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 251–258.

[14] Munjal Desai, Mikhail Medvedev, Marynel Vázquez, Sean McSheehy, Sofia Gadea-Omelchenko, Christian Bruggeman, Aaron Steinfeld, and Holly Yanco. 2012. Effects of changing reliability on trust of robot systems. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 73–80.

[15] Daniel Ellsberg. 1961. Risk, ambiguity, and the Savage axioms. *The quarterly journal of economics* (1961), 643–669.

[16] Craig R Fox and Amos Tversky. 2015. Ambiguity attitudes and social interactions: An experimental investigation. *Journal of Risk and Uncertainty* 38 (2015), 133–150.

[17] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *2007 International symposium on collaborative technologies and systems*. IEEE, 106–114.

[18] Aisling Gallagher, Joanna J Bryson, and Robert H Wortham. 2024. Trust in AI: A systematic review of trust influencing factors and evaluation methods. *Comput. Surveys* 56, 3 (2024), 1–35.

[19] Yue Guo, Xiaowei Yang, Yikun Zhu, Jianyu Yang, and Jing Zhang. 2020. Modeling trust dynamics in human robot teaming: A Bayesian inference approach. *International Journal of Social Robotics* 12 (2020), 665–678.

[20] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[21] Peter A Hancock, Theresa T Kessler, Alexandra D Kaplan, John C Brill, and James L Szalma. 2021. Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors* 63, 7 (2021), 1196–1229.

[22] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[23] Mark Hoogendoorn, S Waqar Jaffry, Peter-Paul van Maanen, and Jan Treur. 2022. Computational models of human trust in different types of automated systems. *Applied Intelligence* 52 (2022), 2693–2711.

[24] Wan-Lin Hu, Kumar Akash, Neera Jain, and Tahira Reid. 2016. Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine* 49, 32 (2016), 48–53.

[25] G Ioanna, M Gianni, MARCO PALOMINO, and G Masala. 2023. I am Robot, Your Health Adviser for Older Adults: Do You Trust My Advice? (2023).

[26] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[27] Catholijn M Jonker and Jan Treur. 2001. A formal analysis of the dynamics of trust based on experiences. *Multiple Approaches to Intelligent Systems* (2001), 221–231.

[28] Kornelia Lazanyi and Greta Maraczi. 2017. Dispositional trust—Do we trust autonomous cars?. In *2017 IEEE 15th International Symposium on Intelligent Systems and Informatics (SISY)*. IEEE, 000135–000140.

[29] Chen Li, Uyanga Turmunkh, and Peter P Wakker. 2019. Trust as a decision under ambiguity. *Experimental Economics* 22, 1 (2019), 51–75.

[30] Jamy Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. *International Journal of Human-Computer Studies* 77 (2015), 23–37.

[31] Bertram F Malle and Daniel Ullman. 2021. A multidimensional conception and measure of human-robot trust. In *Trust in human-robot interaction*. Elsevier, 3–25.

[32] Roger C Mayer, James H Davis, and F David Schoorman. 1995. An integrative model of organizational trust. *Academy of management review* 20, 3 (1995), 709–734.

[33] Linda Miller, Johannes Kraus, Franziska Babel, and Martin Baumann. 2021. More Than a Feeling—Interrelation of Trust Layers in Human-Robot Interaction and the Role of User Dispositions and State Anxiety. *Frontiers in psychology* 12 (2021), 378.

[34] Zahra Rezaei, S Reza Ahmadzadeh, and Paul Robinette. 2024. How humans trust robots: A systematic literature review. *ACM Transactions on Human-Robot Interaction* 13, 1 (2024), 1–36.

[35] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (2016), 101–108.

[36] Alessandra Rossi, Kerstin Dautenhahn, Kheng Lee Koay, Michael L Walters, and Patrick Holthaus. 2020. Evaluating people's perceptions of trust in a robot in a repeated interactions study. In *International Conference on Social Robotics*. Springer, 453–465.

[37] Maha Salem, Gabriella Lakatos, Farshid Amirabdollahian, and Kerstin Dautenhahn. 2015. Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 1–8.

[38] Julian Sanchez, Wendy A Rogers, Arthur D Fisk, and Ericka Rovira. 2014. Understanding reliance on automation: effects of error type, error distribution, age and experience. *Theoretical issues in ergonomics science* 15, 2 (2014), 134–160.

[39] Tracy Sanders, Alexandra Kaplan, Ryan Koch, Michael Schwartz, and Peter A Hancock. 2019. The relationship between trust and use choice in human-robot interaction. *Human factors* 61, 4 (2019), 614–626.

[40] Kristin Schaefer. 2013. The perception and measurement of human-robot trust. (2013).

[41] Kristin E Schaefer, Jessie YC Chen, James L Szalma, and Peter A Hancock. 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human factors* 58, 3 (2016), 377–400.

[42] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. 2020. Multi-task trust transfer for human–robot interaction. *The International Journal of Robotics Research* 39, 2-3 (2020), 233–249.

[43] Charlene K Stokes, Joseph B Lyons, Kenneth Littlejohn, Joseph Natarian, Ellen Case, and Nicholas Speranza. 2010. Accounting for the human in cyberspace: Effects of mood on trust in automation. In *2010 International Symposium on Collaborative Technologies and Systems*. IEEE, 180–187.

[44] Anouk van Maris, Hagen Lehmann, Lorenzo Natale, and Beata Grzyb. 2017. The influence of a robot's embodiment on trust: A longitudinal study. In *Proceedings of the Companion of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. 313–314.

[45] Jin Xu and Ayanna Howard. 2020. How much do you trust your self-driving car? exploring human-robot trust in high-risk scenarios. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 4273–4280.

[46] Rosemarie E Yagoda and Douglas J Gillan. 2012. You want me to trust a ROBOT? The development of a human–robot interaction trust scale. *International Journal of Social Robotics* 4 (2012), 235–248.

[47] X Jessie Yang, Christopher Schemanske, and Christine Searle. 2023. Toward quantifying trust dynamics: How people adjust their trust after moment-to-moment interaction with automation. *Human Factors* 65, 5 (2023), 862–878.

[48] Yan Zhou, Keiko Aoki, and Kenju Akai. 2024. Relationship between health behavior compliance and prospect theory-based risk preferences during a pandemic of COVID-19. *China Economic Review* 86 (2024), 102181.