A Three-Layered Framework for Estimating Human Trust in Robots During Repeated Interactions

Abdullah Alzahrani [1,2*], Simon Robinson [1] and Muneeb Ahmad [1]

[1*] Computer Science Department, Swansea University, Swansea, United Kingdom.
[2] Computer Science Department, Al-Baha University, Al-Baha, Saudi Arabia.

Contributing authors: 2043528@swansea.ac.uk; s.n.w.robinson@swansea.ac.uk; m.i.ahmad@swansea.ac.uk; 2043528@swansea.ac.uk; s.n.w.robinson@swansea.ac.uk; m.i.ahmad@swansea.ac.uk;

Abstract

As robotic systems become more autonomous and capable, they are expected to work alongside humans as teammates rather than just tools. Trust is a crucial factor in collaborative human-robot interaction (HRI), and appropriate trust in robotic collaborators can influence the overall performance of the interaction. Building upon previous work in modelling trust in HRI, this paper describes a refined mathematical trust model to imitate a three-layered framework of trust, which can estimate human trust in robots in real-time. We show that the refined mathematical model significantly outperformed the existing model. Further, this model was tested and validated in a user study where participants engaged with the NAO robot in four sequential collaborative sessions. The results showed that the model is valid based on the linear regression analysis, with both the trust perception score (TPS) and interaction session being significant predictors for the trust modelled score (TMS) computed by applying the trust model. We also demonstrated that trust levels differed across the three layers of trust. This trust model highlights the model's potential in developing adaptive robotic behaviours optimized for user trust, which can enhance the development of robotics systems that can respond to changes in human trust level in real time.

Keywords: Trust, Measurement, Repeated Interactions, Human-Robot Interaction

1 Introduction

The deployment of robotic systems has seen a notable increase in environments where humans and robots work collaboratively [10]. A robot is defined as a re-programmable system capable of performing a wide range of tasks. This system can be autonomous, semi-autonomous, or controlled, and interacts with humans in various capacities, including social robots, self-driving cars, and other automated systems [11]. Establishing successful Human-Robot Interaction (HRI) can enhance efficiency and increase productivity for both humans and robots [5, 49]. Trust is essential to ensure smooth Human-Robot Collaboration (HRC). However, incorrectly calibrated trust can result in either over-reliance or under-reliance, potentially leading to the disuse of these advanced robotic systems [53]. Recognizing this, researchers in HRI are investigating how to develop an online measurement to sense trust in real-time [37]. However, it presents a challenge to model humans' trust in robots as factors affecting humans' trust in robots vary across different settings [9].

In HRI, there are primarily two methods to measure trust [36]. Subjective methods, which are straightforward and direct, involve gauging people's perception toward robots either before or after an interaction, as exemplified by several studies [31, 43, 55]. Nonetheless, these methods have limitations, particularly when it comes to real-time scenarios where preventing the misuse of robots is paramount. In contrast, objective methods delve into the realm of real-time data, observing the actions and reactions of both humans and robots during an interaction. They estimate trust by evaluating aspects like the robot's performance and error rates, as illustrated by Ahmad et al. [3]. However, to systematically estimate trust, a combined approach may be necessary, taking into account factors like interaction duration and the robot's overall performance, as suggested by Law, Scheutz [40].

Mathematically capturing the essence of trust in HRI is a challenging task [60]. Several studies have attempted to create real-time mathematical models of trust [19, 28, 34, 37]. However, these models are not without limitations. One primary issue is that the validation of these models has occurred in simulated environments, which raises questions about their practicality in real-world HRI scenarios [37]. Another aspect is evaluating trust dynamics in the context of repeated and long-term HRI. While there has been limited exploration into factors that influence trust during repeated and long robot interactions [44, 46], the landscape remains largely unexplored.

Exploring further into this, integrated from Lee, See [42], Hoff, Bashir [25] conceptualised a three-layered model of trust: dispositional, situational, and learned (both initial and dynamic), where dispositional, situational and initial trust reflects humans' trust in robots prior to interaction, and dynamically learned trust reflects trust in robots post-interaction. The previous work has aimed to mathematically model trust dynamics in repeated HRI [2] based on Hoff, Bashir [25]. Ahmad et al. attempted to emulate dynamically learned trust dynamics, and found that the time (interaction session) was a significant predictor of the Trust Modelled Score (TMS), whereas subjective Trust Perception Score (TPS) ratings did not predict the TMS. Additionally, the authors found that perceived risk influenced participants' behaviour, with higher risk leading to increased distrust. Moreover, participants' trust behaviour was affected by their perception of the robot's performance compared to its actual performance.

While the fundamental relationship between performance outcomes and trust is well-established across automation, robotics, and AI/ML domains, our contribution extends beyond this basic principle by developing a comprehensive mathematical framework that integrates performance with other critical factors including risk perception, ambiguity aversion, and user control to estimate trust in real-time during repeated interactions. In this paper, we integrate these factors into a three-layered trust model and validate whether this mathematical framework provides accurate trust estimation during repeated physical HRI. In addition, we design a novel game-based experimental task and validate our Trust Modelled Score (TMS) equation across multiple interaction sessions. We further explore the dynamics of dispositional, situational, and dynamically learned trust layers, providing a quantitative approach that advances beyond descriptive models to predictive, implementable trust estimation for robotic systems. We aim to investigate 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 HRI in a collaborative setting?

RQ2 Given the variations in the correlation among the three dimensions of trust, how does dynamic-learned trust evolve during repeated HRI in a collaborative setting?

RQ3 Does the interplay or correlation among the three dimensions of trust (dispositional, situational, and learned (initial and dynamic) trust) exhibit variation during repeated HRI in a collaborative setting?

RQ4 "Is refined mathematical modelling more accurate than current methods in estimating trust in robots?"

The novel contributions (C) of this paper are:

C1 We present a refined mathematical model of the three layers of trust during cooperative HRI, incorporating factors that affect the experience, including risk and ambiguity aversion, building upon insights and limitations identified in previous work.

C2 We validate the model's efficacy using a novel game-based task and show that subjective ratings of trust perceptions strongly predicted the estimation of trust computed by applying the developed model.

C3 We find strong empirical evidence showing linear relationships between different layers of trust as described by Hoff, Bashir [25] in a collaborative HRI task.

C4 We compared two versions of the model and found a significant difference in predicting TMS. The refined trust model outperformed the initial model.

2 Background

2.1 Trust - Conceptualisation

Trust is crucial for the successful operation of any team [21]. Despite being studied in various disciplines, it is challenging to establish a comprehensive definition. In this paper, we consider the following definition: Abbass et al. [1] defined trust as "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". This definition highlights the evolution of trust through experience and interactions, which is critical in studying long-term HRI and enabling successful collaborations.

In light of this interpretation, trust has been categorised into three types: dispositional, situational and learned [25]. **Dispositional trust** refers to the user's tendency to trust the robot before interaction occurs. Dispositional trust is stable over time and is much more related to the user's cultural background, age, gender, and personality. Studies have shown differences in trust behaviour between people of different cultures, age groups, and personality types [9, 41]. **Situational trust** is based on factors external to the user and related to the interaction environment, including task type, complexity, difficulty, perceived risks, and workload. The other factors are internal to the user, including self-confidence and the user's knowledge and expertise. Studies have shown that these factors can affect human trust [16, 52, 56]. Finally, **Learned trust** is based on the user's overall evaluations and experiences with the robotic system before the first interaction (initial trust). During a new interaction with a robotic system (dynamically learned trust), humans' experience affects their established trust level. Experience significantly influences human trust in robots in HRI [24, 50] and can be influenced by the robot's performance and risk during current or repeated interactions and can influence the trust in future interactions [54]. This paper builds on previous research [2] that demonstrated changes in trust over time, the potential influence of risk, and the disparity between a user's perception of a robot's performance and its actual performance. This study delves deeper into the dynamics of trust by examining all three levels of trust and developing a dynamic model to understand how trust evolves over time. We achieve this by integrating risk and differences between user perception and actual robot performance into the calculation of experience.

2.2 Measuring Humans' Trust in Robots

2.2.1 Assessment Methods and Metrics

Prior work has commonly used subjective methods [43, 55, 58], objective methods [35, 36, 40] and psycho-physiological measurements for trust [4, 8]. Additionally, researchers have also attempted to mathematically model human's trust in robots [19, 23, 27, 29, 39, 51]. For instance, Freedy et al. [19] developed a decision-analytical-based measure of trust and conducted two initial experiments to examine trust in a human-robot collaborative task (a simulation environment called MITPAS). The model classified trust in robots based on the self-confidence demonstrated by humans into three categories: under-trust, proper-trust, or over-trust. Hoogendoorn et al. [27] developed trust models with biased experience. The models have been evaluated against empirical data and have shown the impact of bias in the measurement of trust. Saeidi, Wang [51] utilised the trust and self-confidence model to reduce human cognitive workload and improve the overall performance of the human and the robot. This model was tested and validated through a simulated experiment during HRC, which showed its effectiveness in capturing human behaviour and improving overall performance. Hale et al. [23] developed a trust model that reflects a robot's level of cooperation over time and quantifies the amount of information a robot can gain based on its cooperation. The study used simulations to illustrate the trust-driven privacy framework. The results showed that the model was able to capture trust. When a robot stops contributing to a decrease in the cost, the trust and privacy levels decrease, leading to an increase in the amount of added noise to the human's state. Hu et al. [29] introduced a quantitative trust model to study human trust behaviour in human-machine interactions. They conducted an experiment where participants simulated driving a car with an obstacle detection sensor based on an image-recognition algorithm, deciding to trust the algorithm's report based on prior experience. Results showed that the model accurately captures human trust dynamics during interaction based on past experience.

2.2.2 Trust in Repeated or Long-term interactions

In general, there is limited research focusing on measuring or modelling trust in robots [24, 36] as well as investigating the factors impacting human-robot trust [20, 22, 46, 50, 59] in repeated or long-term interactions. Yanco et al. [59] conducted a study to explore the evolution of trust in automated systems within the automotive and medical domains. They used computer-based surveys distributed through Amazon Mechanical Turk to a wide audience. The study focused on factors such as brand reputation (e.g., Google Car versus a small startup) and scenario criticality (safety-critical versus non-safety-critical) and how these influenced trust. The findings showed that trust levels remained fairly consistent across different survey rounds, indicating that initial trust judgments were predictive of short-term trust stability. However, participants who were more familiar with automated systems expressed lower trust and reported higher perceived workload. While the study offers valuable insights into factors influencing trust in automation, the findings are limited to static, survey-based assessments and may not accurately predict trust in real-time interactions. Hafızoğlu, Sen [22] conducted an experiment to examine the effects of past experiences on trust in repeated interactions with software agents in a collaborative game environment. The study involved participants interacting with virtual agents in a team task setting. The results showed that positive past experiences led to an increase in human trust in their agent teammates, while negative experiences resulted in a decrease in trust. Gremillion et al. [20] developed a model and estimation scheme that can predict changes in decision authority during interactions with a simulated autonomous driving assistant. The study utilized a highly controlled simulated leader-follower driving task, where participants operated a virtual vehicle on a two-lane closed circuit. The vehicle was equipped with an autonomous driving assistant, which could either control only the throttle or both the steering and throttle. Participants had to decide when to toggle driving authority to the autonomous assistant based on the driving conditions while simultaneously performing a secondary task. The primary outcome of the study was the development of models that can predict a driver's trust-based decisions using a range of psychophysiological and environmental data. However, the study was limited to incorporating self-reports from subjects to enhance the model's accuracy. Miller et al. [46] investigated the psychological dynamics of Human-Robot Interaction (HRI), focusing on trust across three layers: dispositional, initial, and learned trust. The study utilized a humanoid robot, TIAGo, a service robot designed for domestic environments. Participants engaged with the robot in a laboratory setting where the robot approached them twice. This interaction was controlled using a Wizard of Oz paradigm, where an operator remotely guided the robot to simulate autonomous behaviour. The study revealed that initial and dynamically learned trust were not significantly associated, suggesting that trust in HRI is dynamic and context-dependent, particularly in tasks requiring close physical proximity to a humanoid robot. Alarcon et al. [7] examined the dynamics of trust before and after trust violations in human-human and human-robot interactions. Participants were paired with either a human or a NAO robot partner in a modified trust game, where trustworthiness was manipulated through violations of ability (performance), benevolence, and integrity. The results showed that participants were less forgiving of performance-based errors from robots, supporting the Perfect Automation Schema (PAS). Moreover, robots were perceived more negatively than humans even in cases of non-performance-based violations, suggesting persistent biases against robots. These findings highlight the asymmetry in trust attribution and the critical role of performance expectations in shaping trust in robots. Rossi et al. [50] evaluated how the timing of errors in repeated interactions with a humanoid companion robot (Care-O-bot 4) influenced human trust. Their study found that the timing of errors in repeated and long-term human-robot interactions, whether at the beginning or end, correlates with a loss of human trust in the robot.

Research on trust in repeated or long-term HRI has revealed several key theoretical insights that collectively inform our understanding of trust dynamics. Studies have consistently demonstrated that trust is dynamic and evolves over time based on interaction experiences [22, 46, 50], with the timing and nature of errors significantly impacting trust development [50]. Furthermore, familiarity with robots can paradoxically lead to lower trust levels as users become more aware of system limitations [15, 59]. Research has also shown that trust attribution differs between human-human and human-robot interactions, with robots being less forgiven for performance-based errors, supporting the Perfect Automation Schema [7].

However, these findings, while valuable, primarily confirm established principles: that trust changes based on performance outcomes and user experiences. The critical theoretical gap lies not in understanding that trust changes, but in developing mathematical frameworks capable of predicting and quantifying these changes in real-time during physical HRI. Most existing studies have been limited to static, survey-based assessments [59], simulated interactions [12, 20], small sample sizes [50], or image-based robot representations [14], raising questions about their applicability to real-world physical human-robot interactions.

Prior research has attempted to address this gap by developing mathematical trust models in repeated interactions [2] by emulating the dynamically learned trust framework of Hoff, Bashir [25]. However, evaluation of these models highlighted significant deficiencies: they operated under simplified assumptions (such as initial trust values set at 0.5) and computed user experience based solely on control decisions and perceived robot performance.

What distinguishes our work is not the demonstration that trust changes over time—this is well-established—but rather the development of a comprehensive mathematical framework that can estimate these changes in real-time during physical HRI. In the enhanced model presented in this paper, we have refined the scope of experience calculation to encompass not only user decisions and robot performance but also critical factors such as risk perception and aversion to ambiguity. More importantly, we introduce the Trust Modelled Score (TMS) equation as a novel mathematical tool that integrates multiple factors within a three-layered trust framework, advancing beyond descriptive models to provide predictive, quantitative trust estimation for robotic systems.

3 Trust Model

The trust model is based on three layers of trust: **dispositional**, **situational** and **dynamically learned** [25]. **Dispositional trust** is a reflection of an individual's built-in trust propensity that remains stable over time [25]. **Situational trust** represents the trust level before interaction that is influenced by factors such as the user's knowledge, self-confidence, task type and perceived risks. **Dynamically learned trust**, represents users' experience over time through iterative interactions, incorporating both dispositional and situational trust.

3.1 Initial Model

The foundation of the initial trust model started focusing on the learned trust (initial and dynamic) [2]. We adopted the experiential learning model [33] to mathematically represent trust dynamics, expressed as:

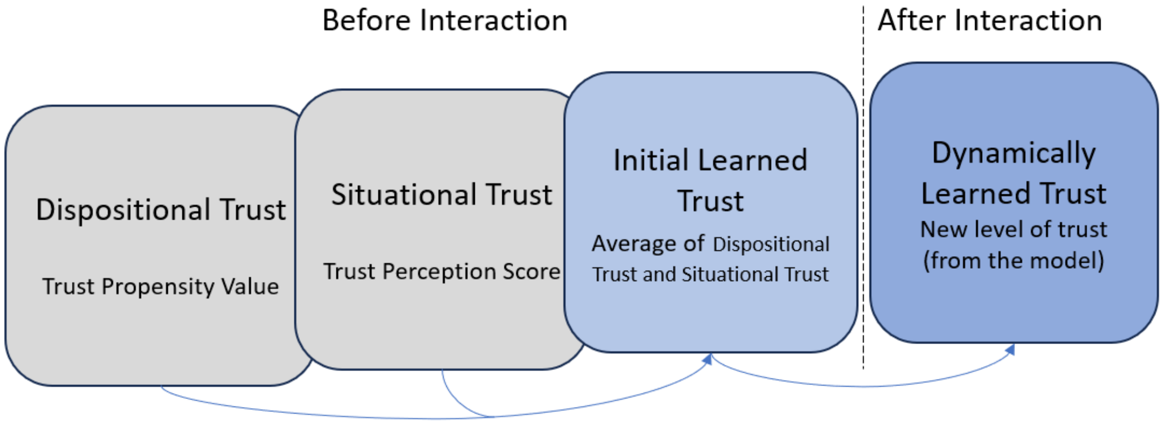

Fig. 1 Modeling the Three Layers of Trust.

$$T(t + \Delta t) = T(t) + \gamma(E(t) - T(t))\Delta t, \quad (1)$$

where $t \geq 0 \subseteq \mathbb{Z}$ represents the count of interaction events, $E(t)$ is the experience and $T(t)$ is the dynamically learned trust at tth interaction, and $T(0)$ is the initial trust at $t = 0$, i.e. when no interactions have occurred. Here, Δt represents the unit difference between events. Thus, $\Delta t = 1$.

The model delineates three distinct cases of trust evolution:

1. Trust increases if the experience $E(t)$ exceeds current trust $T(t)$.
2. Trust remains stable if $E(t)$ equals $T(t)$.
3. Trust declines if $E(t)$ is less than $T(t)$.

The rationale for comparing $E(t)$ to the current trust level $T(t)$ and using it to infer trust at the subsequent time step $T(t + 1)$ is rooted in the understanding that trust at any given moment is not an isolated event. Instead, it is intrinsically linked to the trust levels before and after that instance. As the interaction progresses, each experience $E(t)$ serves as a snapshot of trust, capturing how the user's trust is shaped by the immediate context. This instance of trust then influences the trust level in the next moment, $T(t+1)$, creating a continuous feedback loop where trust dynamically adjusts in response to ongoing experiences.

The idea that experience influences trust is supported by empirical studies in the HRI field. For instance, Miller et al. [46] emphasizes that trust in robots is heavily influenced by prior experiences, particularly in repeated interactions where users can observe and evaluate the robot's performance over time.

The model's central element is the experience, which is calculated based on human decision behaviour and robot performance in a competitive game task:

$$E(t) = \sum_{i=1}^{t} \frac{P_i C_i}{K} \text{ or } 1 \text{ for } K = 0, \quad (2)$$

where P_i and C_i are performance and user control indicators, respectively, and K is the number of taking control.

3.2 Extended Model

Building on the initial model, the extended version further explores the dynamics of trust in HRI. This model is designed to estimate human trust in the trustworthiness of a robot, particularly in situations that present risk and uncertainty. We attempt to model three layers of trust: dispositional, situational and learned (initial and dynamically), as shown in Figure 1.

In this approach, we have chosen specific scales to compute different aspects of trust, aligning with the best practices in trust measurement within

HRI as detailed by Krausman et al. [38]. For computing **dispositional trust (DT)** values, we utilised a Likert scale questionnaire [17]. We computed the **situational trust (ST)** value using the trust perception scale [54]. The initial trust can be better reflected by averaging propensity and situational trust, which considers past pre-interaction experiences with the system. Therefore, we considered the **initial learned trust** $T(0)$ as the average of dispositional and situational trust:

$$T(0) = \frac{DT + ST}{2}. \tag{3}$$

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.

The **dynamically learned trust** is built on the initial model but with differences in the experience computation as:

$$T(t + \Delta t) = T(t) + \gamma(E(t) - T(t))\Delta t, \tag{4}$$

where $t \in \mathbb{N}$ marks the count of interaction events, $E(t)$ is the experience at the complete tth interaction, and $T(t)$ is the dynamically learned trust. Here, $\gamma \in [0, 1]$ is the learning rate, $\gamma = 0.25$, $\Delta t = 1$, represents the unit difference between events.

Based on the definition provided earlier, we can identify the following scenarios:

$$\begin{aligned} &Scenario1 \quad T(t + \Delta t) > T(t);\text{ if } E(t) - T(t) > 0\\ &Scenario2 \quad T(t + \Delta t) = T(t);\text{ if } E(t) - T(t) = 0\\ &Scenario3 \quad T(t + \Delta t) < T(t);\text{ if } E(t) - T(t) < 0 \end{aligned}$$

- **Scenario 1:** Trust in the next interaction $T(t+\Delta t)$ increases if the difference between the user's experience $E(t)$ and the current trust level $T(t)$ is positive.
- **Scenario 2:** Trust remains unchanged $T(t + \Delta t) = T(t)$ if the difference between the experience and the trust level is zero.
- **Scenario 3:** Trust decreases in the subsequent interaction $T(t + \Delta t)$ if this difference is negative.

As the experience $E(t)$ is the key component of the model, we will explore the computation of the experience to extend the model. The rationale for comparing $E(t)$ to the current trust level $T(t)$ and using it to infer trust at the subsequent time step T(t+1) is rooted in the understanding that trust at any given moment is not an isolated event. Instead, it is intrinsically linked to the trust levels before and after that instance. As the interaction progresses, each experience $E(t)$ serves as a snapshot of trust, capturing how the immediate context shapes the user's trust. This instance of trust then influences the trust level in the next moment, $T(t + 1)$, creating a continuous feedback loop where trust dynamically adjusts in response to ongoing experiences. In this version, it is calculated based on human decision-making behaviour, the performance of robots, risk, and ambiguity aversion in a given task as follows:

$$E(t) = (1 - (\frac{\sum_{i=1}^{N} |P_i C_i - C_i R_i|}{N})) - A(t) \tag{5}$$

Where P_i, C_i, and R_i are context-dependent indicators of performance, human control, and risk, respectively, at the ith instance, N is the total number of interactions, and $A(t)$ represents ambiguity aversion. Both P_i and C_i are task-specific and are binary variables with possible values of 0 or 1. The risk R_i is categorized into two fundamental levels: low and high (0,1), respectively.

The part of the equation $|P_i C_i - C_i R_i|$ measures how well the robot's performance aligns with the user's decisions and associated risks over time. This is because the user's actions can be affected by the performance and the risk, making it important to consider both when evaluating the alignment between the robot and the user to assess the experience $E(t)$. Dividing by N normalizes $|P_iC_i - C_iR_i|$, ensuring it remains within a standardized range and providing a consistent measure of alignment between the robot's performance, user's decisions, and associated risks, irrespective of the number of interactions.

Subtracting $(\frac{\sum_{i=1}^{N}|P_iC_i - C_iR_i|}{N})$ from 1 inverts its scale, converting a measure of misalignment into alignment. This is key since $E(t)$ signifies trust, which increases with better alignment between robot performance, user decisions, and risks.

We understand that $E(t)$ can be influenced by the difference between anticipated and actual robot failure rates. We have integrated the concept of ambiguity aversion, represented by $A(t)$, into the model to account for the uncertainties users might face regarding the frequency of robot failures and the potential impact of this uncertainty on user control and experience.

$$A(t) = \frac{\sum_{i=1}^{N}|K_i - F_i|}{N}, \tag{6}$$

Where K_i is the expected number of robot failures (how many times the user overrides the robot), F_i is the actual number of robot failures at time t, and N is the total number of instances. With this representing of $E(t) \in [0, 1] \subset \mathbb{N}$, and an initial $T(0) \in [0, 1] \subset \mathbb{N}$, it is clear that $T(t) \in [0, 1]$ with 1 representing a complete trust, and 0 illustrating a complete distrust; see Figure 2.

4 Study Design

We designed a study to validate the mathematical trust model, involving participants interacting with the NAO robot on four different occasions during collaborative HRI, with each session lasting approximately 7.45 minutes. Each session contained multiple decision points where participants had to decide whether to accept or reject the robot's suggestions. At each decision point, the model computed instantaneous trust, dynamically updating it throughout the session based on these interactions. By the end of each session, the cumulative experience, combined with the previous trust score, formed a new trust level. After each session, participants completed a questionnaire to assess their perceived trust in the robot.

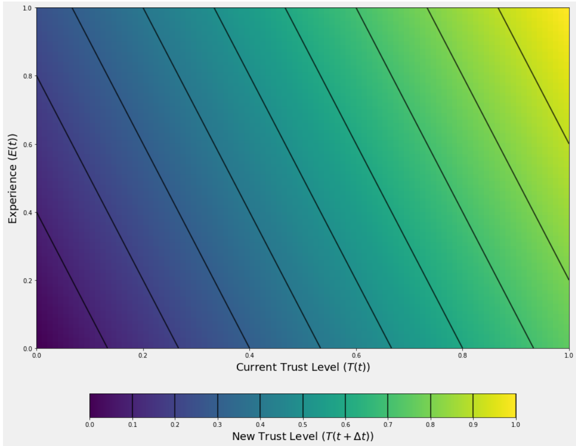

Fig. 2 Illustration of the impact of Current Trust Levels $T(t)$ and Experiences $E(t)$ on the New Trust Level $T(t+\Delta t)$ for $\gamma = 0.25$, showing that a highly positive experience has a limited impact when current trust is low.

This setup allowed us to compare the model's real-time computed trust scores with participants' self-reported trust levels. All participants followed the same sequence of four interactive sessions to ensure consistency in the study conditions. While randomisation is often used in such studies, we chose a fixed session order to focus on measuring trust dynamics over time. This uniform approach allowed us to observe trust evolution consistently across participants. All sessions occurred on the same day, with a 5-minute interval between sessions. We tested the following hypotheses:

H1: Both the Trust Perception Score (TPS) and interaction session (time) will predict the Trust Modelled Score (TMS).

H2: We will observe a significant interaction effect on sessions (session1, session2, session3, and session4) for TMS and TPS scores, reflecting that human dynamically learned trust in robots will change over time during repeated HRI in a collaborative setting.

H3: We will observe variations in the interplay or correlation among the three layers of trust – dispositional, situational, and learned (both initial and dynamic).

H4: The refined model will significantly improve the prediction of TMS compared to the initial model.

4.1 System description

The system presented in Figure 3 consists of two modules. The first module is an interactive card game that generates various situations for participants to either trust or distrust the robot. The second module is a semi-autonomous robot that plays the game with the participants and assists them in making decisions. The model is designed to estimate human trust in the trustworthiness of the robot, particularly in situations that present risk and uncertainty. In the Bluff Game, we focus on key factors that impact trust, such as the robot's accuracy in providing advice, the participant's control in accepting or rejecting the robot's advice, and perceived risk (when the player's cards are more than the opponent's), which is indicated by the proportion of the participant's cards to the opponent's cards. The main objective of the system is to analyze the participants' reactions to situations that involve trust with the robot and how the robot's behaviour over time impacts their decisions to trust it.

4.1.1 The Game

We developed the *Bluff Game*, a Python-based interactive card system that allows participants (forming Player 1 with the robot) to play collaboratively as a team against an adversary agent (Player 2). The game consists of 52 cards, including four sets of each ace, numbers 1-10, jack, queen, and king. The interactive interface provides play and decision buttons (accept and reject), enabling smooth interaction between the players and the game. At the beginning of the game, Player 1 and Player 2 receive 15 cards. The game's goal is for players to eliminate all of the cards before the opponent. Whoever eliminates all their cards first wins the game. *Bluff* is a turn-taking game where Player 1 selects a set of 2-4 cards to discard, and Player 2 decides whether to accept or reject the selected cards. If Player 2 accepts, the turn passes without revealing the cards. If Player 2 challenges the claim and it's found to be true, Player 2 must take the discarded cards; if false, Player 1 takes back the cards. The game aims for either player to eliminate all their cards, updating the card list dynamically after each turn. During the game, at each turn, Player 1 discusses decision-making with the robot on which action to take with Player 2's claim (accept or reject). A message appears asking the participants to start the discussion. The robot provides suggestions on decision-making, advising whether to trust or distrust. These suggestions were based on a predetermined strategy that is consistently applied to all participants in every session. This strategy was part of the Wizard of Oz (WOz) method used (see Figure 4), where the robot operator's decisions were pre-scripted. If the player 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 [30, 57].

The primary risk in the Bluff Game revolves around the possibility of losing the game, representing a challenge to participants' ability to trust the robot's suggestions effectively. While losing does not carry severe real-world consequences, it introduces a competitive element that can influence trust dynamics. Participants who are more competitive or motivated to win might perceive the stakes as higher, impacting their decision-making and trust calibration. In scenarios with more significant real-world consequences, such as financial stakes, trust dynamics would likely shift significantly. However, due to ethical considerations and to avoid unnecessary stress on participants, the controlled environment of the Bluff Game allows us to observe trust behaviours ethically while maintaining a balance in perceived risk levels.

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 player has significantly more cards than their opponent. 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. This aspect is crucial for reflecting the complexity and unpredictability present in HRC, effectively simulating real-world scenarios where decisions must be made with incomplete information.

Our selection of the Bluff Game was guided by the fundamental requirement that trust research must involve situations of uncertainty where participants must rely on an agent despite incomplete

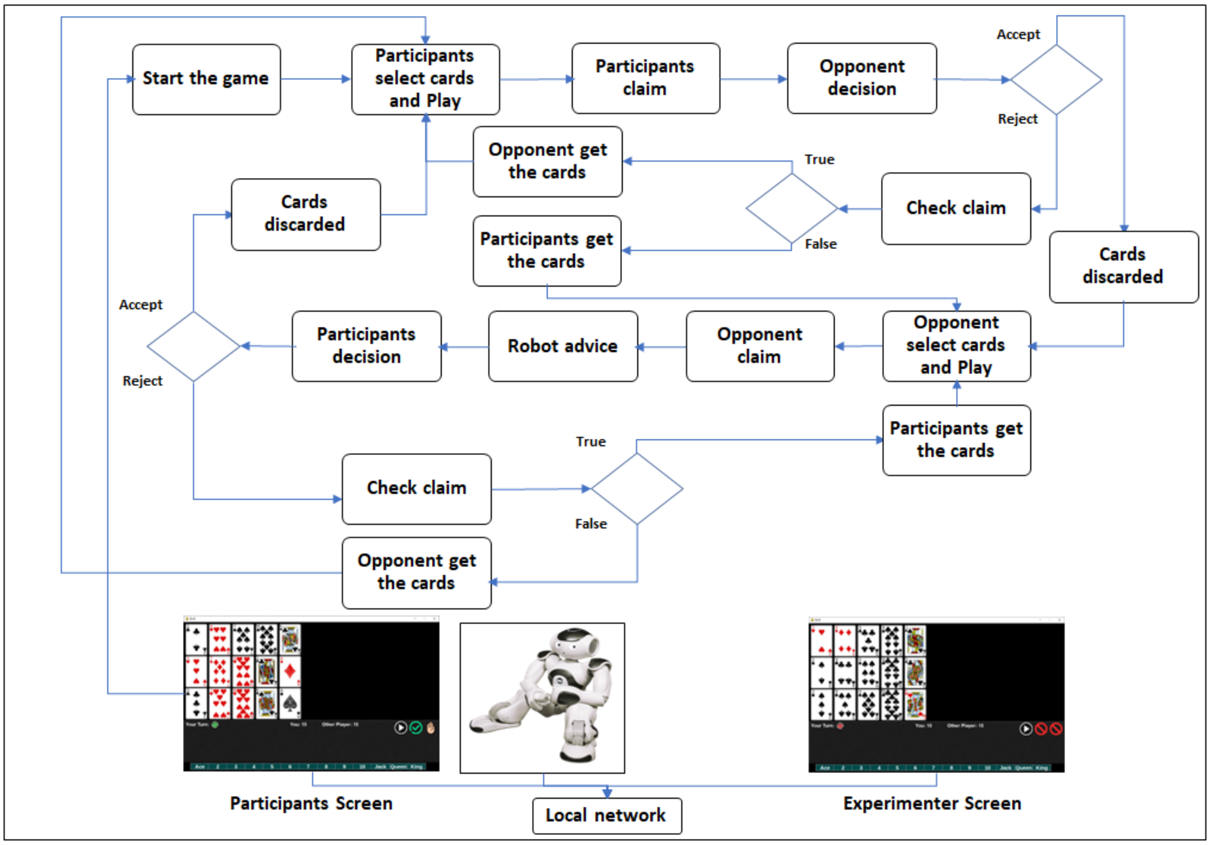

Fig. 3 System overview

information [45]. This approach aligns with established trust research methodologies that employ uncertainty-based tasks to create conditions where trust decisions become meaningful [12, 15, 20].

The calculation of experience $E(t)$ and the dynamically learned trust in the game setting hinges on several key variables. Risk, which can be defined in HRI as an individual's perception of the possible negative consequences associated with interacting with robots [47]. This perception is based on their knowledge and experience of the task, regardless of their personal history or familiarity with the system, technology or person that may be involved in that situation [47]. In this context, Risk was quantified as the risk index R_i. Specifically, R_i is given a value of 1 if Player 2 has more cards than Player 1, which directly impacts the perceived likelihood of negative outcomes (losing the game) if unable to eliminate their cards first. Otherwise, R_i is assigned a value of 0.

The performance P_i equates to 1 when the robot's advice is accurate or when the user controls the incorrect robot's advice. otherwise, $P_i = 0$. Another variable, 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. Our decision to represent these factors as either 0 or 1 was primarily driven by the specific setup of our study, where the interactions and decision-making moments were relatively straightforward. For example, trust decisions often involve clear-cut scenarios, such as whether the robot's advice is accurate or not. In our context, risk is assessed by comparing the number of remaining cards between the participant and the opponent. These variables, along with the Ambiguity Aversion $A(t)$, were essential in computing the experience $E(t)$ and dynamically learned trust during the game.

The term $|P_iC_i - C_iR_i|$ will represent 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 outcome of following the robot's advice was not immediately clear or predictable. For example, the robot might suggest accepting the opponent's claim, but if that claim turned out to be false, the cards would be discarded without revealing their true value. Ambiguity aversion was applied in the following manner: $A(t)$ reflects the user's aversion to uncertainty surrounding the robot's performance. A difference between K_i and F_i in each instance indicates a mismatch between the expected and actual robot performance, contributing to the overall Ambiguity Aversion $A(t)$. This metric is important to understand the influence of the user's uncertainty on their instantaneous trust (experience) in the robot during the game. (see Table 2).

P_i	C_i	R_i	$\lvert P_iC_i - C_iR_i \rvert$
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

Table 1 Truth Table for $|P_iC_i - C_iR_i|$

K_i	F_i	$\lvert K_i - F_i \rvert$
0	0	0
1	0	1
0	1	1
1	1	0

Table 2 Truth Table for $|K_i - F_i|$

4.1.2 Interaction Scenarios

We programmed the Nao robot to interact verbally with participants based on various game events. The game was controlled using the WOz method, and participants were kept uninformed about it to avoid any bias. The interaction was divided into three phases: welcome and introduction to the game, gameplay, and ending of the game.

On the first occasion, the robot welcomed the participants by saying, "Hello. I am a Nao robot. Today, I will assist you in making decisions to "accept" or "reject" in the card game. " and "Now, please get ready and start the game" respectively. Participants engaged in the game on four different occasions. On the second, third, and fourth occasions, the robot greeted the participants and reintroduced them to the game by saying, "Hello again. Thank you for playing; please remember I am here to assist you in deciding to "accept" or "reject". Let's have fun" and "Now, please get ready and start the game" respectively.

Once the game began, the Nao robot informed the participants by saying "The game starts now". Following the game rules, the robot interacted with the participants during various game events. The game's flow involved the robot interacting with the participants during decisions and other situations in the game as follows:

1. During the experiment, the robot consistently followed a predefined protocol and strategy when participants asked about the decision-making process in the accept condition. The robot provided feedback as follows: "Given the game has just started, I think we could accept the claim for now; what do you think?", "I think we could accept, what do you think?", "I suggest accept, what do you think?", or "I think it seems reasonable to accept the claim, what do you think?".
2. In the reject claims condition, the robot said, "I think they might want to discard non-similar cards first, what do you think?", "I think they are bluffing, what do you think?", "I suggest rejecting the claim; what do you think?"
3. If the participants agreed with the robot's suggestion to accept, the robot said "Okay, let's

continue", "Okay, let's proceed", or "Okay, let's see how to conclude".
4. If the participants agreed with the robot's suggestion of rejecting the claims, the robot said "Okay, let's see".
5. If the participants disagreed with the robot's suggestion, the robot said "Okay, it is up to you".
6. If the participants asked the robot to repeat the suggestion, the robot repeated the suggestion for them.
7. If the robot did not hear the participants, the robot said "Sorry, I did not hear that, could you please repeat it".
8. If the participants seem to have been occupied with something else, the robot said "You seem occupied with something else, could you please focus on the game".
9. If the participants asked the robot for anything else during the game, the robot said "I can only advise you when you are deciding to accept or reject".

The robot congratulated or encouraged the participants for the next round at each game's end. If the participants won the game, the robot expressed: "Congratulations on your win! Good luck in the next round". If the participants lost the game, the robot said: "Hard luck, good luck in the upcoming rounds". In the final session, the robot said goodbye and hoped to interact with you soon to its message, announcing the end of the experiment.

4.2 Participants

This study was conducted with 45 participants aged between 18 and 40 years. The age distribution averaged 33.13 years with a standard deviation of 6.22. Out of the 45 participants, 19 were females, 25 were males, and one participant chose not to disclose their gender. We invited participants to partake in the study via the university's electronic mailing system and flyers around the university campus. Participants were able to book their slots for the study using the online scheduling platform *Calendly*[1].

We chose a sample size of 45 participants based on a priori power analysis to ensure sufficient power for detecting significant effects in the study. We conducted the power analysis using G*Power, which indicated that to achieve 80% power for detecting a large effect at a significance level of $\alpha = .05$, a minimum sample size of 43 participants was required for a linear multiple regression test with 2 predictors. Our results showed that R^2 is .750, resulting in a large effect size f^2 of 3.0.

Our sample size of 45 participants is consistent with the norms in HRI research. According to a study by Zimmerman et al. [61], most in-person HRI studies involve fewer than 50 participants. This suggests that our sample size is well within the typical range for studies in this field, providing a solid basis for our findings while still acknowledging the need for larger-scale studies in the future.

To determine the participants' prior interactions with robots, we classified them into four tiers: extensive, moderate, minimal, and none. Those with a background in robot construction or operation were considered to have extensive experience, while individuals who frequently used robots were classified as moderate. Those who had sporadic interactions with robots were labelled as having minimal experience. The breakdown of participants revealed that 11 had extensive experience, 4 had moderate experience, 22 had minimal experience, and 8 had never interacted with robots.

4.3 Setup and Materials

In the study, we utilised two separate rooms, as illustrated in Figure 4. In the first room, the participants had a laptop to play the game while the robot was positioned on the table next to them. The participants were seated beside the robot. The participants used a tablet to complete questionnaires before and after each game round. In the second room, the experimenter sat in front of a laptop to control the game, robot, and overall interaction.

We used the humanoid Nao robot developed by Aldebaran Robotics. Nao is 58cm in height, equipped with an inertial sensor, two cameras, eyes, eight full-colour RGB LEDs, and many other sensors.

[1]https://calendly.com

Fig. 4 Experiment Setup. An experimenter controls the robot in one room (left), while the participant is playing the game with the assistance of the robot in another room (right).

4.4 Procedure

The study was conducted in the following steps:

1. On entering the lab, participants were greeted by the researcher and completed the propensity to trust questionnaire before proceeding with the study.
2. Participants received the experiment information sheet and game instruction sheet and signed the consent form.
3. Participants completed the demographics questionnaire, including information about their experience with the robot.
4. Participants were given a demonstration of the game and had time to practice, allowing for a better understanding of the game and the interaction with the robot.
5. Participants completed the pre-interaction questionnaire.
6. Participants wore glasses and a wristband, and the experimenter began recording the data to be collected from these devices and left the room.
7. Participants engaged in the game alongside the Nao robot, with their interactions being recorded, while the experimenter remotely controlled the gameplay and robot from the other room.
8. After each game, the experimenter walked into the room, asked the participants to complete the post-interaction questionnaire.
9. The rest of the study repeated steps 6, 7, and 8 on three different occasions.
10. At the end, participants were thanked for their participation and were told that they would receive a £10 Amazon voucher as a token of appreciation for their participation in the study.

4.5 Measurements

To accurately assess trust in HRI, we implemented a comprehensive approach, including questionnaires and empirical data that included observations of user control, robot performance, risk and ambiguity aversion. The data was applied to this model, enabling us to calculate TMS.

- Before participating in any interaction or gaining awareness of the surrounding environment of the interaction, the participants were asked to complete a 10-item questionnaire on the tablet to assess dispositional trust [17]. This questionnaire utilised a 5-point Likert scale ranging from "Strongly Agree" to "Strongly Disagree" for responses. The items on the questionnaire are detailed in Table 3. This scale was recently developed specifically for HRI contexts through the Delphi method with expert input, the scale offers strong content validity and a balanced consideration of both trust and distrust. It reflects the understanding, supported by the Computers Are Social Actors (CASA) paradigm, that human robot trust shares psychological foundations with interpersonal trust [24, 48, 55]. Dispositional trust refers to an individual's general tendency to trust others, shaped by personality and previous experiences [25], and this general measure captures that foundational trait without requiring robot-specific items, making it suitable for assessing trust in HRI contexts.
- After becoming aware of the interaction and the role of the robot, but before the primary interaction, participants completed a pre-interaction questionnaire to assess their situational trust towards the robot by rating the robot on the TPS scale from 0 to 100. The scale consists of 40 items and a subscale of 14 items). The items on the scale are detailed in Table 4. In this study, we utilised the 14-item subscale since it helps measure pre-interaction trust and changes in trust over time and during multiple trials. Following [55], we determined the trust score

by first reverse coding the "have errors," "unresponsive," and "malfunction" items, and then computing the average of all 14 items.

- To validate the model's credibility, we employed TPS subjective measures of trust created by Schaefer [55]. Participants were asked to rate the robot's performance in the game using the aforementioned TPS scale.

5 Results

5.1 H1: Predicting TMS with TPS and Session

To test **H1**, we conducted a multiple linear regression to predict the Trust Modelled Score (TMS) using two main predictors: **Trust Perception Score (TPS)** and **Session (time)**. The **TPS** is a subjective score reflecting participants' perception of trust in the robot during different stages of interaction, while the **Session** represents the time points or phases during the experiment in which trust was assessed.

The regression model was found to be highly significant, $F(2, 177) = 265.605, p < .001$, with $R^2 = 0.750$ (Adjusted $R^2 = 0.747$), meaning that 75% of the variance in TMS is explained by TPS and Session Variables (see Figure 5). Both TPS and Session Variables were significant predictors of TMS:

- **TPS**: $b = 0.902, t(177) = 19.986, p < .001$, indicating a strong positive relationship between perceived trust and the modelled trust score.
- **Session**: $b = 0.015, t(177) = 4.825, p < .001$, indicating a significant change in trust across the interactive sessions.

Additionally, a significant correlation was found between TPS and TMS, $r = 0.847, p < .001$, emphasizing the close relationship between participants' subjective trust and the trust predicted by the model (see Figure 6).

To ensure that these findings were robust to the repeated-measures structure of the data, a supplementary mixed-effects regression model with random intercepts for session was also conducted. TPS remained a significant positive predictor of TMS ($\beta = 0.133, SE = 0.058, p = .021$), and Session also remained a significant predictor ($\beta = 0.028, SE = 0.012, p = .016$). The random intercept variance approached zero, indicating that very little unexplained session-level variability remained once TPS and Session were included as fixed effects. The fixed-effects estimates were highly consistent with the linear regression, confirming that the conclusions for H1 are robust.

5.2 H2: The Effect of Interactive Sessions on TPS and TMS

To test **H2**, a repeated-measures ANOVA was conducted to examine the effect of interactive sessions on TPS and TMS. The analysis demonstrated significant variation in TPS and TMS across the four interactive sessions:

- **TPS**: $F(3, 42) = 6.994, p < .001$
- **TMS**: $F(3, 42) = 15.917, p < .001$

Post hoc pairwise comparisons (using Bonferroni correction) showed the following results:

- For **TPS**, there was a significant increase between session 1 and session 3 ($p = 0.026$) and between session 1 and session 4 ($p = 0.007$), while no significant differences were observed between sessions 2 and 3 or sessions 3 and 4.
- For **TMS**, significant increases were found between session 1 and each subsequent session: session 2 ($p < .001$), session 3 ($p < .001$), and session 4 ($p < .001$). No significant differences were observed between sessions 2 and 3 or sessions 3 and 4.

The mean and standard deviations for TPS and TMS across sessions are presented in Table 5.

5.3 Hierarchical Regression Analysis

To validate the unique contribution of our Trust Modelled Score (TMS) beyond temporal effects, we conducted hierarchical regression analyses predicting Trust Perception Scores (TPS).

In the first step, TMS was entered as a predictor of subjective trust ratings, accounting for 6.3% of the variance, $R^2 = .063$, $F(1, 182) = 12.28$, $p < .001$.

Item No.	Statement
1	I suspect hidden motives in others.
2	I am suspicious of other people's intentions.
3	You can't be too careful in dealing with people.
4	It is better to be cautious with strangers until they have shown they are trustworthy.
5	I feel that other people can be relied upon to do what they say they will do.
6	Most people are honest in their dealings with others.
7	I generally give people the benefit of the doubt when I first meet them.
8	I generally trust other people unless they give me a reason not to.
9	I trust what people say.
10	Trusting another person is not difficult for me.

Response	**Strongly Agree**	**Agree**	**Neutral**	**Disagree**	**Strongly Disagree**

Table 3 Dispositional Trust Questionnaire Items

Item No.	Statement
1	Dependable.
2	Reliable.
3	Predictable.
4	Act consistently.
5	Function successfully.
6	Meet the needs of the mission/task.
7	Provide appropriate information.
8	Communicate with people.
9	Provide feedback.
10	Follow directions.
11	Perform exactly as instructed.
12	Have errors.
13	Unresponsive.
14	Malfunction

Response	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**

Table 4 Trust Perception Scale (TPS) 14-item Subscale

In the second step, interaction session was added as an additional predictor, explaining a further 4.0% of the variance, $\Delta R^2 = .040$, $F(1, 181) = 8.16$, $p = .005$. This resulted in a total R^2 of .104.

These results demonstrate that our mathematical trust model significantly predicts subjective trust perceptions and explains unique variance beyond temporal effects alone, thereby providing empirical validation for the utility of the TMS equation in estimating trust in human–robot interaction (HRI).

5.4 H3: Differences Across Trust Layers

To test **H3**, a repeated-measures ANOVA was used to explore the differences in human trust across the **dispositional**, **situational**, and **dynamically learned trust** layers. The results showed significant differences between these trust layers, $F(5, 40) = 58.907, p < .001$.

We conducted Pearson correlation tests to assess the relationships between the different trust layers:

- **Dispositional trust (DT)** and **Situational trust (ST)** showed a significant positive correlation ($r(43) = 0.309, p = 0.039$).
- **Situational trust (ST)** and **Dynamically learned trust (LT)** were also positively correlated ($r(43) = 0.536, p < .001$).
- **Dispositional trust (DT)** and **Learned trust (LT)**, represented by objective TMS measurements, were significantly correlated ($r(43) = 0.563, p < .001$).

5.5 H4: Comparison of Initial and Refined Trust Models

To test **H4**, we compared the initial trust model and the refined trust model, both applied to the data collected during the experiment. A regression analysis was performed for each model to estimate the **Trust Modelled Score (TMS)**. The results showed:

- **Initial model**: $F(2, 177) = 16.066, p < .001$, $R^2 = 0.154$, Adjusted $R^2 = 0.144$.
- **Refined model**: $F(2, 177) = 265.605, p < .001$, $R^2 = 0.750$, Adjusted $R^2 = 0.747$.

To statistically compare the two models, we conducted a one-way ANCOVA, which revealed a significant difference between the models, $F(1, 357) = 18.893, p < .001$. The **refined model** demonstrated a stronger predictive capability, as indicated by the higher R^2 value, showing improved fit and predictive power compared to the initial model (Figure 7).

	TPS		TMS	
Session	Mean	SD	Mean	SD
1	.8027	.1322	.6236	.0727
2	.8324	.1163	.6702	.0626
3	.8469	.1035	.6841	.0628
4	.8522	.1183	.6910	.0980

Table 5 Means and Standard Deviations (SD) for TPS and TMS across Sessions

6 Discussion

This study investigated modelling human trust in robots during repeated collaborative HRI. In this section, we discuss how our empirical findings link

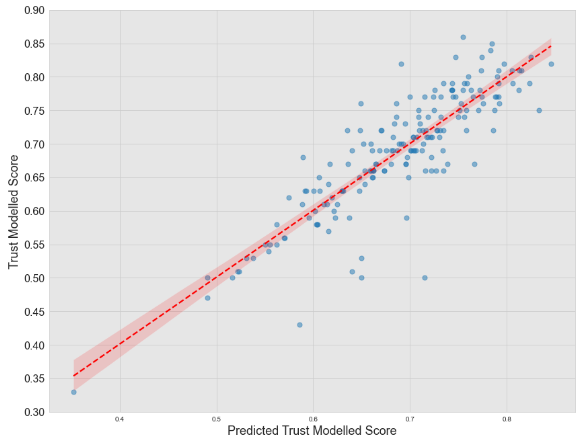

Fig. 5 A regression plot displaying the relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and session variables.

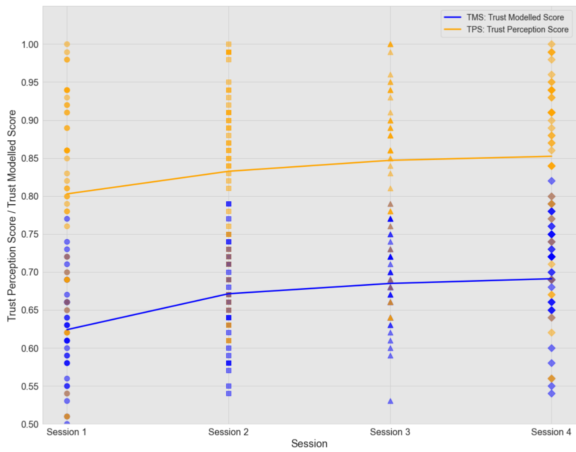

Fig. 6 Scatter plot depicting the changes in the trust perception score (in Orange) and trust modelled score (in Blue) over time.

back to established trust theories and, crucially, how our refined mathematical model expands the existing theoretical knowledge of trust.

6.1 Predicting Trust Modelled Score (TMS)

Our findings confirmed H1, demonstrating that both the Trust Perception Score (TPS) and the interaction session (time) are significant predictors for the Trust Modelled Score (TMS), which was computed from our model. This marks a notable

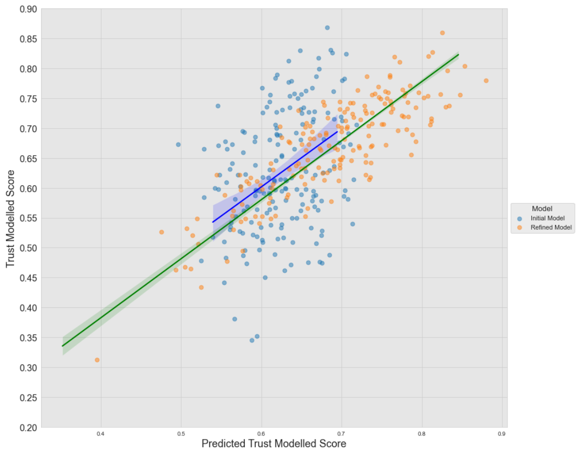

Fig. 7 Comparison of regression lines for the initial (blue) and refined (green) trust models, illustrating the improved predictive capability of the refined model in estimating the TMS.

expansion of prior work [2] where TPS did not emerge as a significant predictor. The enhanced predictive power in this refined model is a direct theoretical advancement stemming from our deliberate integration of risk and ambiguity aversion into the calculation of the user's experience, $E(t)$.

Theoretically, human trust is not simply a function of a system's objective performance but is deeply intertwined with psychological factors such as the perceived risks involved and one's comfort with uncertainty. By mathematically formalising how these factors influence the "experience" that feeds into trust updates, our model provides a more granular and psychologically informed understanding of trust formation and dynamics. This addresses RQ1, as it demonstrates how the model, by incorporating these nuanced elements that contribute to dispositional, situational, and learned trust, more accurately captures and accounts for their interplay in real-time HRI, thus expanding our theoretical grasp of multi-layered trust dynamics.

6.2 Evolution of Dynamic-Learned Trust Over Time

Regarding H2, our results showed that both TPS and TMS changed significantly over time across the four interactive sessions. This acceptance of H2 directly addresses RQ2, examining how dynamic-learned trust evolves in a collaborative HRI setting. These findings strongly align with experiential learning theories of trust [6, 32], which posit that trust is a dynamic construct continuously updated by ongoing interactions. Our work empirically substantiates this by showing quantifiable shifts in both perceived and modelled trust across sessions.

Further enriching this theoretical understanding, our analysis of the contributing factors revealed significant, dynamic changes in risk perception across sessions. The observed decrease in perceived risk from session 2 onwards suggests that as participants gained familiarity and experience with the robot's capabilities within the task, their assessment of potential negative outcomes shifted. This highlights that it is not just the occurrence of experiences, but the recalibration of contextual factors like risk based on these experiences, that drives trust evolution. Whilst ambiguity aversion did not show significant session-to-session differences in this specific game context, its inclusion in the experience calculation still contributes to a more complete instantaneous trust update, reflecting the user's comfort with uncertainty. This demonstrates how our model offers a mechanism to theoretically explain and quantify how specific elements within the "experience" feedback loop contribute to the evolving nature of dynamically learned trust.

6.3 Interplay Among Trust Layers

Our study confirmed H3, revealing variations and correlations among the three distinct layers of trust – dispositional, situational, and learned (initial and dynamic) – during HRC. This directly answers RQ3, providing empirical evidence for the relationships proposed in theoretical frameworks like that of Hoff, Bashir [25].

We observed a significant positive correlation between dispositional trust (DT), representing an individual's general propensity to trust others [17], and situational trust (ST), assessed after participants were introduced to the specific experimental task [55]. This empirically supports the theoretical notion that a fundamental, inherent trust propensity can indeed influence an individual's initial trust judgement in a novel HRI context. Whilst some studies, such as Driggs, Vangsness [18], have found inverse relationships depending on task difficulty, our findings suggest that in a collaborative and moderately challenging game, a baseline willingness to trust carries over to the initial assessment of a robotic partner.

Furthermore, we found that situational trust was positively correlated with dynamically learned trust (LT). This empirically established link is crucial as it demonstrates a continuous influence from the initial contextual assessment of the robot to the trust that develops through repeated interaction. This finding provides a nuanced perspective, as it contrasts with some prior research [46] that suggested a potential disconnect between initial and learned trust. Our results imply that in tasks involving sustained collaboration and cumulative experience, the initial situational assessment remains a relevant anchor for subsequent trust development. The positive correlation between dispositional trust and learned trust (TMS) further reinforces the theoretical idea that an individual's fundamental trust orientation can continue to exert an influence on how trust accumulates and evolves over extended interactions. These empirical correlations collectively demonstrate that the distinct layers of trust are interconnected, and their interplay is modulated by the specific task and interaction dynamics, offering a more robust and empirically grounded understanding of Hoff, Bashir [25]'s framework.

6.4 Superiority of the Refined Trust Model and Expansion of Knowledge

The acceptance of H4 demonstrates that our refined model significantly improved the prediction of TMS compared to the initial model, directly addressing RQ4. This substantial increase in the refined model's predictive power is not merely a statistical improvement but signifies a profound theoretical and practical advancement in our understanding and modelling of HRI trust.

The key to this enhanced performance, and indeed its contribution to the body of knowledge, lies precisely in the new equation for the experience component, $E(t)$, within our mathematical framework. Previous computational models of trust in HRI often relied on simpler feedback loops, perhaps solely based on whether the robot's action was "correct" or "incorrect" relative to a performance metric. Our refined $E(t)$ equation (Equation 5), particularly through its integrated terms, expands our knowledge of trust by formalising how crucial psychological nuances are integrated into its real-time computation.

The term relating to the alignment of performance, human control, and risk (derived from the first part of Equation 5) moves beyond a binary success/failure. It mathematically captures the idea that a user's experience and subsequent trust update are influenced not just by the robot's accuracy, P_i, or the user's decision, C_i, but critically, by how these align with the perceived risk, R_i, of the situation. This formalises the theoretical understanding that trust is context-dependent and risk-sensitive; a correct action in a low-risk scenario might build less trust than an equally correct action in a high-risk scenario where the robot's reliability is truly put to the test. This provides a computational mechanism for how risk directly mediates the impact of performance on trust, a crucial refinement over simpler performance-based models.

The inclusion of ambiguity aversion, $A(t)$, as defined in Equation 6 and derived from the discrepancy between expected, K_i, and actual, F_i, robot failures, is another significant theoretical expansion. Trust theory recognises that uncertainty (ambiguity) about a system's reliability can inhibit trust, even if performance is generally good. Our equation provides a concrete, mathematical way to integrate this psychological factor, showing how a user's aversion to unpredictable robot behaviour (or unexpected failures) directly modulates the overall "experience" that feeds into the trust model. This moves beyond simply reacting to observed failures and accounts for the user's mental model and expectations of robot fallibility, thus offering a more complete picture of trust dynamics under uncertainty.

In essence, this new equation for $E(t)$ allows the model to become a more psychologically valid and comprehensive computational model of trust. It provides a concrete, quantitative mechanism for understanding how these nuanced factors – risk perception, ambiguity, and their interaction with performance and user control – mathematically combine to update trust in real-time. This is a significant leap from descriptive trust models, offering a predictive, quantitative, and implementable framework that aligns more closely with the multifaceted complexities of trust in HRI as described by Hoff, Bashir [25].

6.5 General Implications and Future Work

The findings of this study have important implications for both HRI research and the broader theory of trust. Firstly, our results strongly suggest that trust in robots is not a static attribute but a dynamic construct that evolves over time, heavily influenced by repeated interactions and changing contextual factors. This underscores the critical need for more long-term studies in HRI to fully capture the nuances of trust development and decay. Secondly, the enhanced predictive power of our model, achieved through the explicit incorporation of psychological factors like risk and ambiguity aversion, highlights their theoretical significance in shaping human trust. This provides a more comprehensive understanding of trust and offers a pathway for designing more truly trustworthy and context-aware robotic systems. Lastly, the observed correlations between dispositional, situational, and learned trust layers suggest that fundamental principles from social psychology and interpersonal trust theories remain highly relevant and valuable for refining trust models for HRI.

A key limitation of this study is that whilst we captured participants' prior experience with robots, we did not conduct a direct correlation analysis between this experience and the various trust measures. Prior experience is a well-documented factor influencing trust in automation and robotics, as individuals with more exposure often calibrate their trust differently. Future studies should incorporate such an analysis to explore how familiarity with technology affects trust development over time.

Additionally, our participant pool was primarily composed of university students. Whilst this demographic offers advantages such as greater familiarity and comfort with new technologies [26] and well-developed cognitive abilities for complex tasks [13], it may also introduce biases. University students might approach interactions with a more critical mindset, potentially scrutinising robot performance more rigorously than individuals in non-academic environments. This could lead to different trust dynamics compared to populations where trust might be more readily given, such as those in care settings, or where educational backgrounds and prior technology exposure vary. Therefore, future research should explore trust dynamics across more diverse participant groups to enhance the generalisability of our findings.

Similarly, other binary variables in our model, such as performance ($\mathrm{P_i}$) and control ($\mathrm{C_i}$), could benefit from continuous representations in future iterations. For instance, performance could be measured on a scale reflecting degrees of success rather than simple success/failure, and control could represent the degree of intervention rather than a binary choice to trust or not trust.

A further limitation of this study is the absence of significant real-world risk in the experimental design. While the *Bluff Game* was designed to create both a collaborative and competitive environment that introduced a level of uncertainty, it did not involve any monetary or other high-stakes risks for the participants. The concept of trust is intrinsically linked to the presence of risk, and the lack of significant consequences for poor decisions may have influenced the participants´trust behaviours. Future research should aim to incorporate more substantial risks, such as financial incentives or penalties, to create a more ecologically valid environment for studying human-robot trust.

7 Conclusion & Future Work

In this paper, we built upon prior work to present a refined mathematical model that emulates the three-layered (initial, situational, learned) trust framework and potentially estimates human trust in robots in real-time during repeated HRI. The findings confirmed the model's validity, with both TPS and the sessions being significant predictors for the TMS. Notably, the refined model demonstrated a significant improvement in predicting the TMS more effectively than the initial model. This increase in performance validates the enhancements made to the model, highlighting its increased precision in trust estimation. The validation of this model can be attributed to several enhancements. We integrated additional task-dependent factors, such as risk and ambiguity aversion, which significantly refined the model's ability to shape the experience. Testing the model in different contexts further highlighted its adaptability and robustness, demonstrating its improved capability to assess human trust in real-time. The implications of having a validated trust measurement are substantial. This model opens up opportunities for a variety of applications, such as reinforcement learning, where the model can help in shaping reward functions. Consequently, this facilitates the development of behaviours in robotic systems that optimise user trust across various tasks, thereby enhancing the effectiveness and adaptability of HRI.

In the future, we will primarily focus on applying this validated model's capabilities within the reinforcement learning domain to develop adaptive robotic systems that can optimise human-robot trust. Also, we will undertake further validation testing and refinement of the model to enhance its adaptability, accuracy, and applicability across diverse HRI contexts.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethical Statement The study was submitted for ethical review and was approved by the university ethics board. Approval number: 2202370516013.

Data availability The datasets analysed during the current study are available from the corresponding author on reasonable request

References

[1] *Abbass Hussein A, Scholz Jason, Reid Darryn J.* Foundations of trusted autonomy. 2018.

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

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

[4] *Ajenaghughrure Ighoyota Ben, Sousa Sonia C, Kosunen Ilkka Johannes, Lamas David.* Predictive model to assess user trust: a psycho-physiological approach // Proceedings of the 10th Indian conference on human-computer interaction. 2019. 1–10.

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

[6] *Alaieri Fahad, Vellino André.* Ethical Decision Making in Robots: Autonomy, Trust and Responsibility: Autonomy Trust and Responsibility // Social Robotics: 8th International Conference, ICSR 2016, Kansas City, MO, USA, November 1-3, 2016 Proceedings 8. 2016. 159–168.

[7] *Alarcon Gene M, Capiola August, Hamdan Izz Aldin, Lee Michael A, Jessup Sarah A.* Differential biases in human-human versus human-robot interactions // Applied Ergonomics. 2023. 106. 103858.

[8] *Alzahrani Abdullah, Ahmad Muneeb.* Crucial Clues: Investigating Psychophysiological Behaviors for Measuring Trust in Human-Robot Interaction. In INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23), October 09–13, 2023, Paris, France. ACM, New York, NY, USA // In INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '23). 2023.

[9] *Alzahrani Abdullah, Robinson Simon, Ahmad Muneeb.* Exploring Factors Affecting User Trust Across Different Human-Robot Interaction Settings and Cultures // Proceedings of the 10th International Conference on Human-Agent Interaction. New York, NY, USA: Association for Computing Machinery,

2022. 123–131. (HAI '22).

[10] *Barfield Jessica K.* Self-Disclosure of Personal Information, Robot Appearance, and Robot Trustworthiness // 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN). 2021. 67–72.

[11] *Chen Ang, Yin Ruixue, Cao Lin, Yuan Chenwang, Ding HK, Zhang WJ.* Soft robotics: Definition and research issues // 2017 24th international conference on mechatronics and machine vision in practice (M2VIP). 2017. 366–370.

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

[13] *Dautenhahn Kerstin.* Methodology & themes of human-robot interaction: A growing research field // International Journal of Advanced Robotic Systems. 2007. 4, 1. 15.

[14] *De Visser Ewart J, Peeters Marieke MM, Jung Malte F, Kohn Spencer, Shaw Tyler H, Pak Richard, Neerincx Mark A.* Towards a theory of longitudinal trust calibration in human–robot teams // International journal of social robotics. 2020. 12, 2. 459–478.

[15] *Desai Munjal.* Modeling trust to improve human-robot interaction. 2012.

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

[17] *Dragostinov Yavor, Harðardóttir Daney, McKenna Peter Edward, Robb David A, Nesset Birthe, Ahmad Muneeb Imtiaz, Romeo Marta, Lim Mei Yii, Yu Chuang, Jang Youngkyoon, others .* Preliminary psychometric scale development using the mixed methods Delphi technique // Methods in Psychology. 2022. 7. 100103.

[18] *Driggs Jade, Vangsness Lisa.* Changes in Trust in Automation (TIA) After Performing a Visual Search Task with an Automated System // 2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS). 2022. 1–6.

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

[20] *Gremillion Gregory M, Donavanik Daniel, Neubauer Catherine E, Brody Justin D, Schaefer Kristin E.* Estimating human state from simulated assisted driving with stochastic filtering techniques // Advances in Human Factors in Simulation and Modeling: Proceedings of the AHFE 2018 International Conferences on Human Factors and Simulation and Digital Human Modeling and Applied Optimization, Held on July 21–25, 2018, in Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9. 2019. 113–125.

[21] *Groom Victoria, Nass Clifford.* Can robots be teammates?: Benchmarks in human–robot teams // Interaction studies. 2007. 8, 3. 483–500.

[22] *Hafızoğlu Feyza Merve, Sen Sandip.* The effects of past experience on trust in repeated human-agent teamwork // Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems. 2018. 514–522.

[23] *Hale Matthew T, Setter Tina, Fregene Kingsley.* Trust-Driven Privacy in Human-Robot Interactions // 2019 American Control Conference (ACC). 2019. 5234–5239.

[24] *Hancock Peter A, Kessler Theresa T, Kaplan Alexandra D, Brill John C, Szalma*

James L. Evolving trust in robots: specification through sequential and comparative meta-analyses // Human factors. 2021. 63, 7. 1196–1229.

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

[26] *Hoffman Guy, Zhao Xuan.* A primer for conducting experiments in human–robot interaction // ACM Transactions on Human-Robot Interaction (THRI). 2020. 10, 1. 1–31.

[27] *Hoogendoorn Mark, Jaffry S Waqar, Maanen Peter-Paul van, Treur Jan.* Modeling and validation of biased human trust // 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology. 2. 2011. 256–263.

[28] *Hu Wan-Lin, Akash Kumar, Reid Tahira, Jain Neera.* Computational modeling of the dynamics of human trust during human–machine interactions // IEEE Transactions on Human-Machine Systems. 2018. 49, 6. 485–497.

[29] *Hu Wan-Lin, Akash Kumar, Reid Tahira, Jain Neera.* Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions // IEEE Transactions on Human-Machine Systems. 2019. 49, 6. 485–497.

[30] *Ioanna G., Gianni M., Palomino Marco, Masala G.* I am Robot, Your Health Adviser for Older Adults: Do You Trust My Advice? // Journal or Conference Name. 2023.

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

[32] *Jonker Catholijn M, Schalken Joost JP, Theeuwes Jan, Treur Jan.* Human experiments in trust dynamics // Trust Management: Second International Conference, iTrust 2004, Oxford, UK, March 29-April 1, 2004. Proceedings 2. 2004. 206–220.

[33] *Jonker Catholijn M, Treur Jan.* Formal analysis of models for the dynamics of trust based on experiences // Multi-Agent System Engineering: 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99 Valencia, Spain, June 30–July 2, 1999 Proceedings 9. 1999. 221–231.

[34] *Kaniarasu Poornima, Steinfeld Aaron, Desai Munjal, Yanco Holly.* Potential measures for detecting trust changes // 2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2012. 241–242.

[35] *Kaniarasu Poornima, Steinfeld Aaron, Desai Munjal, Yanco Holly.* Robot confidence and trust alignment // 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2013. 155–156.

[36] *Khavas Zahra Rezaei.* A review on trust in human-robot interaction // arXiv preprint arXiv:2105.10045. 2021.

[37] *Khavas Zahra Rezaei, Ahmadzadeh S Reza, Robinette Paul.* Modeling trust in human-robot interaction: A survey // International Conference on Social Robotics. 2020. 529–541.

[38] *Krausman Andrea, Neubauer Catherine, Forster Daniel, Lakhmani Shan, Baker Anthony L, Fitzhugh Sean M, Gremillion Gregory, Wright Julia L, Metcalfe Jason S, Schaefer Kristin E.* Trust measurement in human-autonomy teams: Development of a conceptual toolkit // ACM Transactions on Human-Robot Interaction (THRI). 2022. 11, 3. 1–58.

[39] *Kumar Bimal, Dubey Akash Dutt.* Evaluation of trust in robots: A cognitive approach // 2017 International Conference on Computer Communication and Informatics (ICCCI). 2017. 1–6.

[40] *Law Theresa, Scheutz Matthias.* Trust:

Recent concepts and evaluations in human-robot interaction // Trust in human-robot interaction. 2021. 27–57.

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

[42] *Lee John D, See Katrina A.* Trust in automation: Designing for appropriate reliance // Human factors. 2004. 46, 1. 50–80.

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

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

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

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

[47] *Rachel E. Stuck Brianna J. Tomlinson, Walker Bruce N.* The importance of incorporating risk into human-automation trust // Theoretical Issues in Ergonomics Science. 2022. 23, 4. 500–516.

[48] *Reeves Byron, Nass Clifford.* The media equation: How people treat computers, television, and new media like real people // Cambridge, UK. 1996. 10, 10. 19–36.

[49] *Robert Lionel P., Dennis Alan R., Ahuja Manju K.* Differences are Different: Examining the Effects of Communication Media on the Impacts of Racial and Gender Diversity in Decision-Making Teams // Information Systems Research. 2018. 29, 3. 525–545.

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

[51] *Saeidi Hamed, Wang Y.* Trust and self-confidence based autonomy allocation for robotic systems // 2015 54th IEEE Conference on Decision and Control (CDC). 2015. 6052–6057.

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

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

[54] *Schaefer Kristin E.* The perception and measurement of human-robot trust // Journal or Conference Name. 2013.

[55] *Schaefer Kristin E.* Measuring trust in human robot interactions: Development of the "trust perception scale-HRI" // Robust intelligence and trust in autonomous systems. 2016. 191–218.

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

[57] *Xu Jin, Howard Ayanna.* How much do you trust your self-driving car? exploring

human-robot trust in high-risk scenarios // 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). 2020. 4273–4280.

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

[59] *Yanco Holly A, Desai Munjal, Drury Jill L, Steinfeld Aaron.* Methods for developing trust models for intelligent systems // Robust intelligence and trust in autonomous systems. 2016. 219–254.

[60] *Zahedi Zahra, Verma Mudit, Sreedharan Sarath, Kambhampati Subbarao.* Trust-aware planning: Modeling trust evolution in longitudinal human-robot interaction // arXiv preprint arXiv:2105.01220. 2021.

[61] *Zimmerman Megan, Bagchi Shelly, Marvel Jeremy, Nguyen Vinh.* An analysis of metrics and methods in research from human-robot interaction conferences, 2015–2021 // 2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI). 2022. 644–648.