Modelling Human Trust in Robots During Repeated Interactions

Muneeb Imtiaz Ahmad
Swansea University
Swansea, UK
m.i.ahmad@swansea.ac.uk

Abdullah Alzahrani
Swansea University
Swansea, UK
2043528@swansea.ac.uk

Simon Robinson
Swansea University
Swansea, UK
s.n.w.robinson@swansea.ac.uk

Alma Rahat
Swansea University
Swansea, UK
a.a.m.rahat@swansea.ac.uk

ABSTRACT

Modelling humans' trust in robots is critical during human-robot interaction (HRI) to avoid under- or over-reliance on robots. Currently, it is challenging to calibrate trust in real-time. Consequently, we see limited work on calibrating humans' trust in robots in HRI. In this paper we describe a mathematical model that attempts to emulate the three-layered (initial, situational, learned) framework of trust capable of potentially estimating humans' trust in robots in real-time. We evaluated the trust model in an experimental setup that involved participants playing a trust game on four occasions. We validate the model based on linear regression analysis that showed that the trust perception score (TPS) and interaction session predicted the trust modelled score (TMS) computed by applying the trust model. We also show that TPS and TMS did not change significantly from the second to the fourth session. However, TPS and TMS captured in the last session increased significantly from the first session. The described work is an initial effort to model three layers of humans' trust in robot in a repeated HRI setup and requires further testing and extension to improve its robustness across settings.

CCS CONCEPTS

• **Human-centered computing → Human Robot interaction** ; User studies; • **Computer systems organization** → Robotics.

KEYWORDS

Trust, Measurement, Repeated Interactions, Human-Robot Interaction

ACM Reference Format:
Muneeb Imtiaz Ahmad, Abdullah Alzahrani, Simon Robinson, and Alma Rahat. 2023. Modelling Human Trust in Robots During Repeated Interactions. In *International Conference on Human-Agent Interaction (HAI '23), December 4–7, 2023, Gothenburg, Sweden.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3623809.3623892

HAI '23, December 4–7, 2023, Gothenburg, Sweden
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0824-4/23/12…$15.00
https://doi.org/10.1145/3623809.3623892

1 INTRODUCTION

Trust is a significant factor to achieve smooth human-robot interaction (HRI) in both competitive and collaborative settings [25, 29]. Recently we have witnessed limited work on human trust during human-robot competitive contexts [29, 50]. However, most of the work has considered collaborative contexts [17]. Competition is an integral aspect of daily life, as it can enhance performance, problem-solving effectiveness, and enjoyment [38]. Robots have increasingly become part of our life, from automated cleaning systems and transportation to personal assistants [20]. As technology advances, robots are being designed and deployed to compete with humans in various domains, such as economics, sports, and games [27, 51]. In competitive settings, the truthfulness of the robot is considered for establishing trust [42]. The existing definition of trust concerning truthfulness indicates that trust is "*generalized expectancy held by an individual that the word, promise, oral or written statement of another individual or group can be relied on*" (p.26) [43]. We understand that robots' truthfulness influence human trust in robots in situations requiring individuals to rely on or trust an advice or information given by the robot in different settings . In this work, we focus on truth-telling and consider it as performance indicator of the robot and humans relying on the information represent their confidence in the robot. We used these factors to mathematically modal human trust in the robot in a repeat HRI.

Typically, researchers use two methods to measure trust during HRI [23]. Subjective methods to measure trust are popular choices due to their ease of use and involve the use of questionnaires [18, 31, 48]. The questionnaires are used to measure trust pre- or post-interaction with the robot. However, subjective methods are not commonly used in real-time, particularly in the context of understanding the disuse of the robotic system, where it remains necessary to determine the appropriate level of trust to adapt to prevent mis- or dis-use. Objective methods of measuring trust analyse user behaviours during an interaction with robots and consequently can be used in real-time. For example, robot performance or error rate can be objectively computed during an interaction [2]. Law and Scheutz [29] has identified four categories of objective measures: 1) task intervention refers to the frequency of humans' intervention in the robots' task, 2) task delegation refers to humans' preferences of a robot in a team of robots, 3) behaviour change refers to analysing human's behaviour and 4) humans' seeking robot advise [23].

Existing work has used these categories of objective measures individually and not in combination. We understand that these categories individually are meaningful as empirical evidence shows they affect human's trust in robots [12]. However, representing these categories of objective measures together will consider taking different factors into account and will present a more robust representation of trust experienced by a user during HRI. It is indeed challenging to represent the concept of trust mathematically in HRI [59]. Efforts have been made to represent it in the past [10, 21, 40]. We see little work in the existing literature on modelling human trust in robots in the collaborative settings. Besides, most of these trust models have been tested and validated in simulated environments, and some may not be relevant across different HRI settings [24]. Lastly, to the best of knowledge, we did not see how truthfulness of the robot can be used to model human trust in robots in the competitive settings.

Another aspect is the utilisation and testing of trust model during repeated or long-term HRI. We witness limited work on studying factors affecting trust during long-term interaction with robots. Few examples include [35, 55]. Factors such as interactive experience become extremely critical when modelling humans' trust over time in robots because humans use knowledge from past interactions to assess the trustworthiness or trustfulness of a robotic system [13, 34, 42]. In this regard, Hoff and Bashir [13] presented a trust model and highlighted three layers of trust: dispositional, situational, and learned (initial & dynamically learned) trust. Research shows changes in situational trust is varied by human's context-dependent traits (self-confidence & expertise), while dynamically learned trust varies based on experiences gained through the evaluation of a system over time [6, 13]. Both situational and learned trust are closely related to each other [32]. We understand that context-dependent factors such as self-confidence and expertise are impacted from the experience gained by interacting with the robot over time and hence does impact the learned trust in the robot [12].

Considering these aspects, we aim to delve into the following research questions: **RQ1:** How can we individually and as a combination mathematically model the concept of situational and dynamically learned trust during HRI?
RQ2: How does dynamically learned trust evolves with time during repeated HRI?
RQ3: What is the relationship between different layers of dynamically trust during HRI?

Addressing these questions, the novel contributions of this paper are as follows:

(1) We propose a mathematical model for measuring trust, and validate its efficacy through a long-term HRI task.
(2) Using a questionnaire [48], we show that the model predictions of trust scores strongly agree with trust perception of participants in an experiment, where they interacted with a NAO robot in a *novel trust based game.*
(3) We show that the dynamically learned trust varies significantly over time. However, the significant differences generally occur at the last interaction with the robot.
(4) We show that there exist a strong positive correlation between different layers of dynamically learned trust.

The rest of the paper is structured as follows. We provide background material in Section 2. The proposed trust model is presented in Section 3. In section 4, we provide a thorough description of the study. The results are discussed in Section 5, and relevant discussions are provided in Section 6. Finally, we draw our conclusions in Section 7.

2 BACKGROUND

2.1 Trust - Conceptualisation

Trust is a multidimensional concept and currently there remains a gap in the literature to establish its definition [1, 12]. We consider the definition suggesting that "*trust is generalised expectancy held by an individual that the word, promise, oral or written statement of another individual or group can be relied on*" (p.26) [43]. More clearly, we understand that *generalised expectancy* is formed based on the set of individual experiences one has with another individual or technology.

In the light of this definition, we consider the work of Hoff and Bashir [14] and attempts to mathematically model the layers of trust. Hoff and Bashir [14] presented a framework for conceptualising trust that has been widely reported in the HRI literature [35, 46, 49]. They categorised trust into three layers: dispositional trust, situational trust, and learned trust. **Dispositional trust** refers to a human propensity to trust robots based on biological and environmental factors. Dispositional trust, unlike a situational and learned trust, is characterised as a relatively stable trait over time. The factors influencing dispositional trust include culture, age, gender, and personality. **Situational trust** measures the construct related to trust dynamics during HRI in a certain environment. Environment-related and user-related factors can influence situational trust. Environment-related factors include task type, complexity, difficulty, perceived risks, and workload. Users' differences can affect trust in robots during HRI. For instance, users with low self-confidence in their ability to perform a task are more likely to trust the robot. Similarly, a user's expertise or familiarity with a subject matter can impact trust [9, 45]. The mental well-being of a human is also essential. Humans in a pleasant mood are likelier to have an initial trust in robots [53]. **Learned trust** is based on evaluations of the robotic system prior to interaction with the robot (initial trust) or insight gained from the current interaction (dynamically learned trust). The robot's performance in the current interaction is the most significant factor affecting learned trust. Experience is a significant factor influencing human trust in robots in HRI [12]. Experience in robots can be built based on robot performance in previous interactions with robots in a particular context [47].

In this paper, we model situational and dynamically learned trust by considering context-dependent and performance-related factors affecting trust in HRI. When modeling dynamically learned trust, we consider the inter-relation of factors affecting situational and dynamically learned trust in HRI. We achieve it by integrating context-dependent factors that can affect dynamically learned trust over time.

2.2 Measuring Humans' Trust in Robots

Besides, the subjective [18, 31, 48] and objective methods [23, 29] of measuring trust during HRI, efforts have been made to mathematically model humans' trust in robots during HRI [10, 11, 15, 22, 28, 44, 56] in collaborative contexts. Freedy et al. [10] described a model for human trust in a human-robot collaborative setting. The model categorised trust into three categories (under-, proper- or over-trust) based on the self-confidence demonstrated by the human in the robot. Hoogendoorn et al. [15] attempted to model interaction bias experienced during the interaction and mitigated the effect of bias in the measurement of trust. Kaniarasu et al. [22] calibrated the operator's trust in the robot by computing trust mismatch. They defined it as "the degree of alignment of user's trust with the robot's current reliability". In summary, trust mismatch was based on the sum of control taken during an interaction. Xu and Dudek [56] presented an OPTimo model to measure a human supervisor's degree of trust in a robot. OPTIMo formulates Bayesian belief over human trust based on the robot's performance on the task over time to generate a real-time estimate of the human's trust. Saeidi and Wang [44] presented a trust and self-confidence model and quantified trust as a function of human performance and robot performance to describe the difference between human-to-robot trust and human self-confidence.

Kumar and Dubey [28] developed an objective measure for trust as a product of the capability index and intention index. They quantified the capability index as a product of the expertise (number of tasks performed) and capacity (performance) of the robot. The intention index was quantified as the product of desire (performance of the robot for its desired tasks) and commitment (robots only attempt the possible actions). Hale et al. [11] attempted to model trust based on the robot's level of cooperation over time. The uniqueness of the model was introducing a discounted function to bound by taking robot performance and introducing the level of cooperation needed to attain a desired level of trust over time. In summary, the performance of the robot is the standout feature of most of these models with a combination of human confidence demonstrated in the robot. Besides, these models are designed for collaborative settings. In addition, most of these models considered simulated environments for evaluation and were neither tested nor validated in real HRI settings. Further, our literature review did not find any prior model that considered truthfulness of the robot as a function to model trust in a competitive setting. Nevertheless, we did see a few studies used a competitive task when studying human-robot trust [25, 36, 41, 50]. This demonstrates the novelty of the presented work. Other methods of modelling trust have been through machine learning approaches [16, 39, 59]. Lastly, we also see efforts to model robot trust in humans based on human performance and faults made by humans [30, 40]. Our approach, based on mathematical principles and tested in real-time, competitive interactions, uniquely contributes to current trust research in human-robot interactions.

2.3 Trust in Long-term interactions

To our knowledge, there is limited work on analysing trust during long-term interaction with robots [12, 23]. Besides, we see rare studies on factors affecting trust during a long-term or longitudinal interaction [5, 35, 55]. Miller et al. [35] investigated the interrelationship between three layers of trust (disposition, initial, learned) during HRI setup. The robot was controlled through a Wizard of Oz paradigm and was enabled to drive toward the participant twice. The findings suggested that the initial and dynamically learned trust were not related to each other in a task that involved analysing trust in relation to robot distance from the human. In summary, these studies present valuable findings but are very task-specific and suggest the need for more empirical research on trust during long-term HRI.

In parallel, we see limited work on studying and validating trust models during repeated interactions [7, 8, 56]. Desai [8] pioneered the work on modelling trust during repeated HRI and showed that trust ratings did not vary across the session. Besides, trust was impacted by familiarity with the robot, as participants familiar with the robot trust it less as compared to others. While these findings were intriguing, the sample size was small. More recently, De Visser et al. [7] contributed a method for calibrating trust in a longitudinal HRI. The model considered team settings and integrated human trust and their expectations from the robot into its planning process to build and maintain trust over the interaction horizon. They validated the model in a study that consisted ten interaction rounds. However, the maze-based game task only had an image-based robot representation. More specifically, we did not see interaction with a real robot. In summary, the work from Desai [8] to Xu and Dudek [56] to De Visser et al. [7] is specific to different domains and remarks on the challenge of creating a general model for trust in HRI. However, robot performance and confidence in robot functions are among the top factors to dictate trust in robots.

The work described here considers *experience* with the robot as a contributing factor to informing human trust. We computed experience (generalised expectancy) by taking past performance (truthfulness demonstrated by the robot) and user confidence (relying on the information given by the robot) in real time. Besides, the novelty of the work lies in modeling different layers of trust [13] and validating the model during a repeated interaction that involves interacting with a real robot.

2.4 Trust in Human-robot Competition

Human-robot competition is an interaction in which both robot and human compete against each other in a given task for a given reward [57]. Few studies have considered competition in physical exercise [58], trusting a more competitive robot as their teammate [36, 41] and playing a game together [50]. Sebo et al. [50] investigated how trust is impacted in a competitive task when the robot breaks a promise and later how it heals based on an apology from the robot. Most recently, Kirtay et al. [25] looked at how emotions change to trust in human-robot interactions during a competitive task where the robot's performance varied. The study used a Pepper robot and recorded the physiological signals of the participants. The results demonstrated that participants' trust in the robot increased as the robot's performance enhanced, and participants showed positive emotions such as happiness and contentment. The study also discovered that negative emotions such as frustration and disappointment reduced trust in robots. In summary, trust is rarely investigated in the human-robot competitive settings compared to human-robot collaboration when humans and robots share the same goal [33]. In a competitive setting, trust is impacted by the truthfulness of the robot and in this paper, we have explored this dimension to model trust.

3 TRUST MODEL

In this paper, we consider three layers of trust: **initial**, **situational** and **dynamically learned** [13]. **The initial trust** is based on the prior attitude of the humans involved and the reputation of the system. As such this can be set before any interactions have taken place. Without any available bias, it would be reasonable to set a neutral initial trust level. On the other hand, **situational trust** is based on self-confidence and the perceived performance of the robot. We treat this as the instantaneous *experience* from an interaction. **The dynamically learned** trust may be interpreted as a function of time that utilises experience, i.e. situational trust, through iterative interactions, and thus develops estimations of trust over time.

With these interpretations, inspired by the experiential model proposed in [19], we relate different layers of trust as follows:

$$T(t+\Delta t) = T(t) + \gamma(E(t) - T(t))\Delta t, \tag{1}$$

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, $\Delta t = 1$.

Given the definition above, we observe the following cases:

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

(1) **Case 1:** Trust in the next interaction $T(t+\Delta t)$ increases if the difference between the user experience with the robot $E(t)$ and their current trust level $T(t)$ is positive.
(2) **Case 2:** Trust remains unchanged $T(t+\Delta t) = T(t)$ if the difference between the user experience with the robot $E(t)$ and their current trust level $T(t)$ is zero.
(3) **Case 3:** Trust decreases in the subsequent interaction $T(t+\Delta t)$ if the difference between the user experience with the robot $E(t)$ and their current trust level $T(t)$ is negative.

The key component in this model is the experience. We compute it based on human decision behaviour, risk and robot performance in a competitive game task as follows:

$$E(t) = \begin{cases} \sum_{i=1}^{t} \frac{P_i C_i}{K} & \text{if } K > 0. \\ 1 & \text{if } K = 0. \end{cases} \tag{2}$$

Here, $E(t)$ is the experience after a number of interactions t, $P_i \in \{0, 1\}$ is the perceived performance indicator of the robot at the ith interaction with $C_i \in \{0, 1\}$ is the associated human contradiction indicator, $\gamma \in [0, 1]$ is the learning rate, and K is the number of times the user contradicts. It should be noted that P_i and C_i are game specific, and therefore, the approach towards setting them is context-dependent. We provide details of how we set P_i and C_i for the experiments in this paper in Section 4.2.1.

Here, we can deduce that $E(t) \in [0, 1] \subset \mathbb{R}$ because given K contradictions the sum of product of $P_i C_i$ will never be greater than K. With this, and an initial $T(0) \in [0, 1] \subset \mathbb{R}$, it is clear that $T(t) \in [0, 1]$ with 1 representing a complete trust, and 0 illustrating a complete distrust; see Figure 1. It is, therefore, reasonable to consider an initial trust of $T(0) = 0.5$ which means that the human has neutral trust at time point 0 [15], in the absence of some high-level ancillary information. We use this value of $T(0)$ throughout this paper.

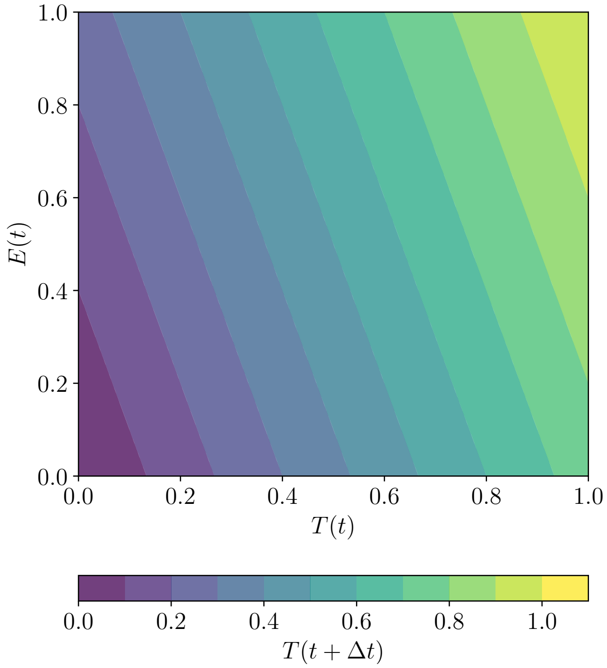

Figure 1: An illustration of how the values of $T(t)$ and $E(t)$ impact $T(t+\Delta t)$ given $\lambda = 0.25$. Unsurprisingly, when trust is low, an immediate highly positive experience does not alter learned trust substantially.

4 STUDY DESIGN

The study was designed to validate the mathematical trust model and involved participants to interact with the NAO robot on four different occasions. All sessions occurred on the same day, with a 5-minute interval between sessions. We tested the following hypotheses:

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

H2: We will observe significant interaction effect on session (session1, session2, session3, and session4) for TMS and TPS scores.

H3: Human dynamically learned trust in robots will change during the repeated interaction.

4.1 Ethics

Since the study involved human participants, an application was submitted to the university ethics board to ensure ethical integrity. The application was approved following a review process. [160322/5031].

4.2 System description

The system shown in Figure 2 consisted of the following modules: 1) a card game inducing situations that enabled participants to either trust or distrust the robot, and 2) a semi-autonomous robot capable of playing the card game with participants. The goal of

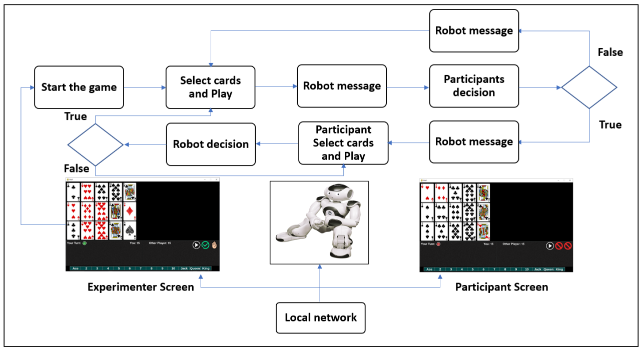

Figure 2: System description

the system was to analyse how participants react in situations involving trusting the robot and how the robot's behaviour over time implicate their trusting decisions in the robot.

4.2.1 The Game. We developed an interactive two-player card game, *Bluff Game*, using python, that participants can play against the robot. The card game consisted of 52 cards with four sets each of ace, 1,2,3,4,5,6,7,8,9,10, jacks, queen and king, a play button, and decision buttons (trust and distrust). Each player gets 15 cards at the start of the game. The goal for each player is to dispose of all the cards before their opponent (another player). Whoever disposes of all their cards first wins the game. It is a turn-taking game. At each turn, a player selects a set consisting of 2-4 cards they intend to dispose of. At this stage, their opponent can either trust or distrust the player on whether they are stating their set of cards correctly that they intend to dispose of or not. For example, if a player states that they have a pair of queens, their opponent will either trust them or distrust them. If the opponent trusts the player, the opponent will take the next turn. The opponent will not be able to view the player's cards, and consequently, cards will be removed from the player's list of cards and opponent will take their turn. Otherwise, when the opponent distrusts the player and asks them to show their cards. In this case, if the player has correctly stated their set of cards, the opponent will receive the players' cards, and consequently, cards will be added to the list of opponent cards. If the player incorrectly states their set of cards, the cards will be returned to the player and the opponent will get their turn. The game continues in the same fashion until one of the players has disposed of all the cards. The game dynamically updates the list of each player's cards at each turn. We conceived the game by considering the factor that it presents situations inducing risk and uncertainty that are in line with the definition of trust. The game puts the player at the risk of losing, where player cards get significantly lesser than the opponent cards.

In this context, considering $B_i \in \{0, 1\}$ as the indicator of whether truly a bluff has occurred or not and $C_i \in \{0, 1\}$ indicating whether the human counterpart has contradicted the robot's claim of the card, we can derive a truth table for the perceived performance of the robot P_i (see Table 1). Using the truth table, we observe that $P_i = 1$ if the user does not contradict the robot's claims irrespective of whether the robot has bluffed or not. On the other hand, when the user contradicts the robot, the value of $P_i \longleftarrow \neg B_i$, i.e. $P_i = 0$ if the robot was truly bluffing and *vice-versa*. We use this in (1) to update trust in the experiments in this paper.

B_i	C_i	P_i
0	0	1
1	0	1
0	1	1
1	1	0

Table 1: Truth table of B_i, C_i and P_i at the ith interaction.

4.2.2 Interaction Scenarios. We programmed the NAO robot to interact verbally with participants during various game events. To prevent bias, we used the Wizard of Oz (WOz) method to control the game without informing the participants. The game comprised two platforms running on separate laptops, with the NAO robot connected via the TCP/IP protocol over a LAN. An experimenter played the game on behalf of the robot and determined whether to bluff based on a predetermined and consistent strategy for all participants. In a separate room, participants played against the robot and made decisions as desired. The interaction involved three phases: a welcome and introduction to the game, playing the game, and ending the game.

The robot welcomed the participant and introduced itself by saying - "Hello. I am a NAO robot. I am going to play a card game against you today. Are you ready?" Participants played the game on four different occasions. On the second, third and fourth occasion, the robot thanked the participant and introduced them to the games by saying - "Hello again. Thank you for playing. We are going to play another game. Are you ready?" and "Let us start" respectively.

Once the game started, the NAO robot informs the participant that "the game starts now". Robot takes the first turn. Following the game rule, the robot interacted with the participant on different game events as follows:

(1) When the robot selected their set of cards, the robot declared them, for example as, "I selected three kings".
(2) When the participant trusted the robot, the robot said: "It is your turn".
(3) When the participant did not trust the robot, and the robot was stating their set of cards correctly, the robot said: "I was telling the truth".
(4) When the participant did not trust the robot, and the robot was not stating their cards correctly, the robot said: "You got me, and it is your turn".
(5) When the robot trusted the participant, "I trust you, and it is my turn."
(6) When the robot did not trust the participant, the robot said: "I think you are bluffing". If the participant was telling the truth, the robot said: "Oh, I was wrong, and it is your turn now".
(7) When the robot did not trust the participant, and the participant was wrong, the robot said "Yes, I got you, and it is my turn now".

At the end of each game, the robot congratulated or wished the participant good luck for the subsequent game. In the winning case, the robot said "Congratulations! You win, thank you and see you in the next round", and in the loose case, it said "You just lost the game, good luck in the following rounds". In the last session, the robot added goodbye to its message, declaring the experiment's end.

4.3 Participants

We recruited 45 participants ranging in age from 18 to 60 (Mean age: 29.77 years, SD = 6.82) 16 identified as female, 28 as male and 1 did not say. Participants were recruited through university mailing lists and flyers around the university campus. The registration for the study was managed using an online application for registration (*Calendly* [1]).

Participants were classified as experienced with robots into high, medium, low and no experience. Participants were categorized as high experienced if they reported having controlled and/or built a robot, medium experienced if they reported using robots several times, and low experienced if they reported interacting with robots on a few occasions. 2 participants had high experience interacting with robots. 2 participants had medium experience interacting with robots. 26 participants had low experience interacting with robots. 15 participants had no experience interacting with robots.

4.4 Setup and Materials

We conducted this study in 2 separate rooms, as shown in Figure 3. In room 1, the laptop was placed on the table for the participant to play the game. The robot was placed across the table in front of the participant. The participant was seated in front of the robot, wearing glasses and a wristband to capture the psychological signals. The participant used a tablet to fill out the demographic and questionnaires after each game round. In room 2, the experimenter was sitting in front of a laptop to control the robot and the 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.

We used E4 Wristband and Pupil Eye Tracking Glasses to capture the physiological data. We recorded blood volume pulse (BVP), heart rate (HR), inter-beat intervals (IBIs) and heart rate variability (HRV) using the wristband. Similarly, we recorded blinking, fixation, and pupil diameter data using the eye tracking glasses. Our objective was to estimate human trust in robots by analyzing the differences in these physiological data during repeated HRI, between trust and distrust states. However, it is not used in the analysis of this paper as it goes beyond the scope of the contributions described in this work.However, we do not report the analysis of the physiological data as this goes beyond the scope of the contributions described in this paper.

4.5 Procedure

The study was conducted in the following steps:

(1) Participants received the experiment information sheet, and game instruction sheet, and signed the consent form.
(2) Participants completed the demographics questionnaire including information about their experience with the robot.

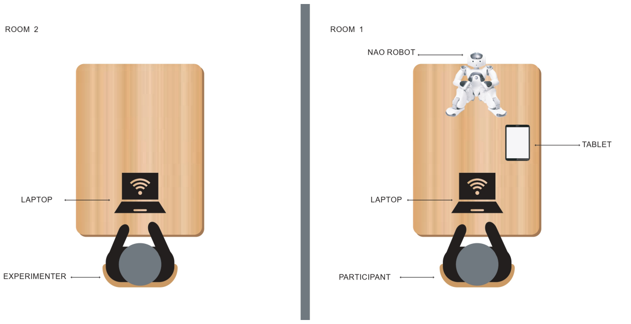

Figure 3: Experiment Setup - it depicts an experimenter controlling the robot in a one room (left), while participant playing the game against the robot in another room (right).

(3) Participants wore glasses and a wristband. The experimenter began the recording of the data to be collected from these devices and left the room.
(4) Experimenter controlled the robot from the other room. Participants played the game against the NAO robot.
(5) After each game, the experimenter walked into the room, stopped collecting the physiological data and asked the participant to complete the questionnaire to rate the robot during the game.
(6) The rest of the study repeated steps 3, 4, and 5 on three different occasions.
(7) At the end, participants were thanked for their participation and were told that they will receive a £10 amazon voucher for their participation in the study.

4.6 Measurements

To measure trust over time during HRI, we collected observed data, including user control and robot performance. We applied the observed data to our model to calculate TMS.

To validate the model, we used TPS subjective measures of trust developed by Schaefer [48]. Participants were asked to rate the robot in the game using a TPS scale, administered using Google Forms on a tablet. The scale has 40 items and a subscale of 14 items, including (function successfully, act consistently, reliable, predictable, dependable, follow directions, meet the needs of the mission, perform exactly as instructed, have errors, provide appropriate information, malfunction, communicate with people, provide feedback, and unresponsive) to rate the robot in percentage. This study used the 14 items subscale because it helps measure changes in trust over time and during multiple trials. Following [48], we calculated the trust score by first reverse coding the 'have errors,' 'unresponsive,' and 'malfunction' items, then computed the average of all 14 items.

We computed the risk during each game turn by dividing the robot's number of cards left by the participant's number of cards left and subtracting them from 1. We assumed that the negative number equals 0, which meant no risk. To compute the risk during the whole game, we calculated the average of each turn during the game. We computed the percentage of the participant's control during the game, which equals the number of times the participant

[1] https://calendly.com

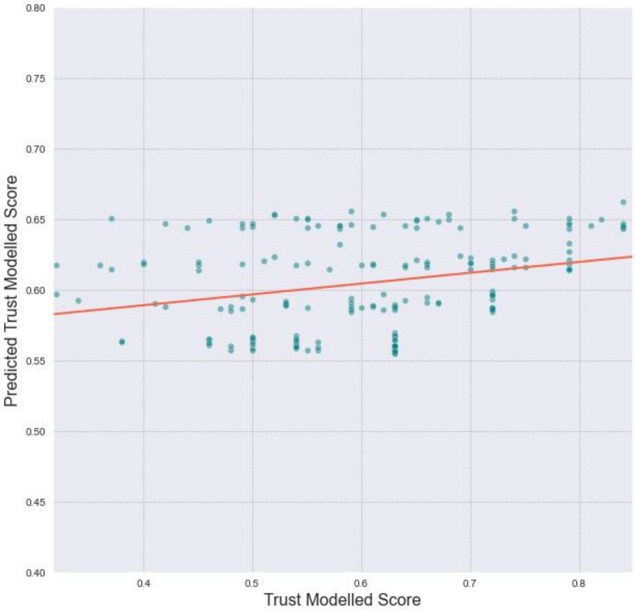

Figure 4: Scatter plot and linear regression line showing a relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and time.

took control divided by the number of turns. We computed the failure rate during the game, which equals the number of robot-perceived failures divided by the number of turns.

SPSS was used for subsequent statistical analyses, providing a comprehensive view of the results and their implications.

5 RESULTS

To test **H1**, a multiple linear regression was calculated to predict TMS based on TPS and session. A significant regression equation was found (F (2,177) = 7.36, $p < .001$), with an R^2 of .077 (see figure 4). Both TPS and TMS increased over time and the session variable was a significant predictor, whereas the TPS variable was not found to be a significant predictor of TMS. Further, we did not witness a correlation between the TPS computed at the end of each session and the TMS computed per game sessions.

To test **H2** and **H3**, a repeated-measures ANOVA was conducted to determine whether there is an effect of interactive session (session 1, session 2, session 3, and seession 4) on TMS and TPS, respectively. We found that both TPS (F(3, 42) =4.08, $p < .01$) and TMS (F(3, 42) = 11.13, $p < .001$) differed significantly across the four interactive session.

A post hoc pairwise comparison using the Bonferroni correction showed a significant increase in both TPS ($p < .02$) and TMS ($p < .001$) between session 1, and session 4 respectively. The increase was not statistically significant for both TPS and TMS in session 2 and session 3 and when comparing session 3 and 4 respectively. The mean and Standard deviation for both TPS and TMS can be seen in Table 2.

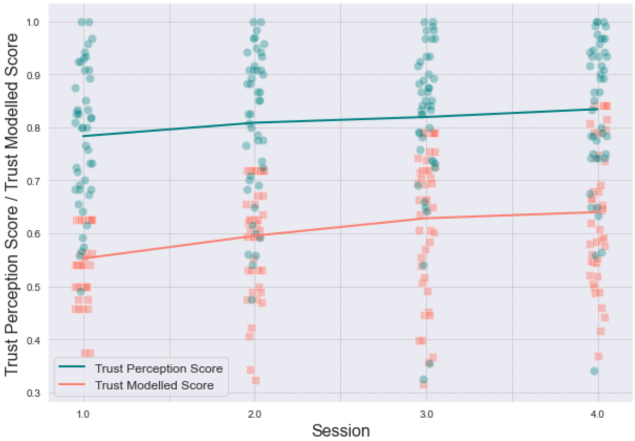

Figure 5: Scatter plot depicting the changes in the trust perception score (in Green) and trust modelled score (in Orange) over time.

Session	N	TMS		TPS	
		Mean	SD	Mean	SD
1	45	0.55	0.07	0.78	0.13
2	45	0.60	0.11	0.81	0.14
3	45	0.63	0.14	0.82	0.15
4	45	0.64	0.13	0.83	0.14

Table 2: Mean (M) and Standard Deviations (SD) of TMS and TPS scores in the 4 sessions

We conducted a Pearson correlation coefficient test to assess the linear relationship between different layers of dynamically learned trust. Both TMS and TPS measurements in each of the four sessions represented layers of learned trust (T_1, T_2, T_3, T_4 respectively). We found that there was a positive correlation between TPS and TMS measured across session ($p < .05$). This means that both TPS and TMS in each session were positively related to each other which suggests a positive increase in dynamically learned trust across the four different sessions.

6 DISCUSSION

H1 indicated that both TPS and time will predict TMS, trust score computed by applying the trust model. The finding showed that the TPS and interaction session significantly predicted TMS. Besides, in Figure 5, we can see how both TPS and TMS show an increase in each interactive session. In particular, the correlation analysis also showed that TPS and TMS in the second, third and fourth sessions were positively related to one other. Hence, **H1** was accepted. We see that TPS alone was not the predictor of TMS in each session and the interaction session was the sole significant predictor of TMS. We see that this is in line with our predictions because TMS was changing based on the experience gained by the participants in each interactive session. In parallel, we understand that the reason that TPS did not independently predict TPS may have been due to many factors involved in the questionnaire as compared to the model [48]. We consider that the model was validated in the holistic view as both measures represented participants' trust in the robot.

We understand that trust in a robotic system changes based on the perceived performance (truthfulness, or error rate) [12, 13, 50]. It is important to note that in the given experimental setup, the perceived performance of the robot remained consistent (around 90% on average) in each of the four sessions. In addition, perceived risk in a given situation impacts trust [26, 37]. Perceived risk refers to an individual's feeling that a specific task or context has potential negative outcomes [54]. In this case, when participants' number of cards left were significantly higher than the robots' number of cards left, it presented a high risk of losing the game. Hence, suggested that participants will take more control. To further understand and explore the effect of these variables on the findings, we computed risk, control/contradiction and failure rate/truthfulness (as described in the section 4.6) experienced during the game on four different occasions. We tested any relationship of risk, contradiction and failure rate/truthfulness with TPS and TMS and between each other across the four sessions. The trends were unique and intriguing as it was only in the third session that perceived risk was positively related to the control. Besides, we did not witness this effect after the third session. It suggests that when perceived risk was high participants significantly contradicted the robot only in the third session. It may be related to the outcome of the previous two games and for most participants, it might have ended in a losing cause. Past studies have shown that a successful or an unsuccessful task outcome impacts user trust [10, 12, 23, 52]. The performance of the robot and participants' contradiction were also positively correlated to each other in the first, second and fourth sessions suggesting the more the robot bluffed, the more participants contradicted the robot. TMS was positively related to risk in the first three sessions, but it was not the case in the last session. These trends across sessions shows that factors affecting trust may impact differently across different situations as shown in the findings of [4]. Lastly, we did not find a relationship between TPS, risk, performance or control in all four sessions.

H2 indicated that there will be an interaction effect on TPS and TMS. The findings confirmed the hypothesis and showed that both TPS and TMS increased significantly over time. We expected this finding because dynamically learned trust changes with experience over time [3, 13, 23]. It was also the case here where trust was changing based on the experience gained during the interactive sessions and experience was computed by analysing participants' confidence in the robot and perceived performance of the robot.

H3 indicated that dynamically learned trust in robots will change with repeated interactions. We found that after the first interaction with the robot both TPS and TMS did not significantly differ between the second, third and fourth sessions. These findings are intriguing and may suggest that once learned, dynamically learned trust does not vary too much. To explore this further, we analysed how risk, performance, and self-confidence (control/contradiction) differed across the four sessions. We found that risk did not differ significantly from the second, third and fourth sessions. Besides, we did not observe significant differences in performance and self-confidence across the four sessions. As it turned out, risk varied between first and all other sessions. Consequently, it was the significant factor impacting trust in all the sessions. In parallel, interesting past literature has shown that with familiarity with the robot in repeat HRI [9], participant seems to trust the robot less. However, our findings were in contradiction. Similar to the familiarity findings, these findings may also be task-centric and be seen respectively.

The work on modelling the three layers of trust by Hoff and Bashir especial in a competitive setting where we used truthfulness as an indicator of trust is novel. But, we note that the presented work is the first effort and understand the complexity. We mainly considered self-confidence and performance to inform situational trust (trust depicted in a given situation). We computed experience based on the game situation. Further, experience in a given interaction session informed the dynamically learned trust over time. We understand from the findings that we can include perceived risk as part of the experience. We will consider the future changes to the model based on this work.

7 CONCLUSION, LIMITATION & FUTURE WORK

In this paper, we presented a mathematical model that emulates the three-layered (initial, situational, learned) framework of trust and can potentially estimate human trust in robots in real-time. The model was evaluated in an experimental setup that involved participants playing a trust game on four different occasions. The model was validated as trust computed based on the model was predicted by the interactive session and the trust perception score measured from the questionnaire. The work draws thought-provoking conclusions. Risk was the sole factor that varied between interactive sessions in line with the trust computed using the model and the trust perception scale questionnaire. This was a fascinating result, and will enable us to integrate it as part of the experience in the future extensions of the model of trust. Further, we found that dynamically learned trust did not change after the first interaction. This finding may be task-centric, and therefore raises the question: how long can it take in a repeated interaction to learn a value of humans' trust in robots that will not vary significantly? Is it a good idea to have learned value of dynamically learned trust, or can this lead to over-trust in robots in a repeated interaction? Lastly, we see confidence (control/contradiction) as a consistent factor in our experiment but understand this also can be task-specific.

Although our contribution has noteworthy novelty and practical implications, some limitations should be considered. Firstly, our model currently operates with a fixed initial trust value set at the natural trust value of 0.5. This may not account for the diverse range of initial trust that individuals may have towards robots. Secondly, our study focused on general trends in trust dynamics but did not consider individual variances in trust trajectories.

In the future, we will also undertake more rigorous testing of the model and will add more factors that may affect trust in robots. This experiment involved a game that presented a competitive setting; so, we aim to test the model in other (cooperative) settings. Lastly, we will reflect on the factor of the number of sessions it can take to learn a value of dynamically learned trust.

REFERENCES

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

[2] Muneeb Imtiaz Ahmad, Jasmin Bernotat, Katrin Lohan, and Friederike Eyssel. 2019. Trust and cognitive load during human-robot interaction. *arXiv preprint arXiv:1909.05160* (2019).

[3] Fahad Alaieri and André Vellino. 2016. Ethical decision making in robots: Autonomy, trust and responsibility. In *International conference on social robotics*. Springer, 159–168.

[4] Abdullah Alzahrani, Simon Robinson, and Muneeb Imtiaz 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 (HAI '22)*. ACM, 121–131.

[5] Franziska Babel, Philipp Hock, Johannes Kraus, and Martin Baumann. 2022. It Will Not Take Long! Longitudinal Effects of Robot Conflict Resolution Strategies on Compliance, Acceptance and Trust. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*. 225–235.

[6] Ana Cristina Costa, C Ashley Fulmer, and Neil R Anderson. 2018. Trust in work teams: An integrative review, multilevel model, and future directions. *Journal of Organizational Behavior* 39, 2 (2018), 169–184.

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

[8] Munjal Desai. 2012. *Modeling trust to improve human-robot interaction*. Ph. D. Dissertation. University of Massachusetts Lowell.

[9] 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 *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 73–80.

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

[17] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, and Rajiv Suman. 2021. Substantial capabilities of robotics in enhancing industry 4.0 implementation. *Cognitive Robotics* 1 (2021), 58–75.

[18] 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.

[19] Catholijn M Jonker and Jan Treur. 1999. Formal analysis of models for the dynamics of trust based on experiences. In *European workshop on modelling autonomous agents in a multi-agent world*. Springer, 221–231.

[20] Takayuki Kanda and Hiroshi Ishiguro. 2017. *Human-robot interaction in social robotics*. CRC Press.

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

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

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

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

[25] Murat Kirtay, Erhan Oztop, Minoru Asada, and Verena V Hafner. 2021. Modeling robot trust based on emergent emotion in an interactive task. In *2021 IEEE International Conference on Development and Learning (ICDL)*. IEEE, 1–8.

[26] Bing Cai Kok and Harold Soh. 2020. Trust in robots: Challenges and opportunities. *Current Robotics Reports* 1, 4 (2020), 297–309.

[27] Alap Kshirsagar, Bnaya Dreyfuss, Guy Ishai, Ori Heffetz, and Guy Hoffman. 2019. Monetary-incentive competition between humans and robots: Experimental results. In *2019 14th acm/ieee international conference on human-robot interaction (hri)*. IEEE, 95–103.

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

[29] Theresa Law and Matthias Scheutz. 2021. Trust: Recent concepts and evaluations in human-robot interaction. *Trust in human-robot interaction* (2021), 27–57.

[30] Harsh Maithani, Juan Antonio Corrales-Ramon, and Youcef Mezouar. 2019. Trust-Based Variable Impedance Control for Cooperative Physical Human-Robot Interaction. In *2019 IEEE International Conference on Mechatronics (ICM)*, Vol. 1. IEEE, 706–711.

[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] Stephen Marsh and Mark R Dibben. 2003. The role of trust in information science and technology. *Annual Review of Information Science and Technology (ARIST)* 37 (2003), 465–98.

[33] Eloise Matheson, Riccardo Minto, Emanuele GG Zampieri, Maurizio Faccio, and Giulio Rosati. 2019. Human–robot collaboration in manufacturing applications: A review. *Robotics* 8, 4 (2019), 100.

[34] John M McNamara, Philip A Stephens, Sasha RX Dall, and Alasdair I Houston. 2009. Evolution of trust and trustworthiness: social awareness favours personality differences. *Proceedings of the Royal Society B: Biological Sciences* 276, 1657 (2009), 605–613.

[35] 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.

[36] Michael Novitzky, Paul Robinette, Michael R Benjamin, Danielle K Gleason, Caileigh Fitzgerald, and Henrik Schmidt. 2018. Preliminary interactions of human-robot trust, cognitive load, and robot intelligence levels in a competitive game. In *Companion of the 2018 ACM/IEEE international conference on human-robot interaction*. 203–204.

[37] LeeAnn Perkins, Janet E Miller, Ali Hashemi, and Gary Burns. 2010. Designing for human-centered systems: Situational risk as a factor of trust in automation. *Proceedings of the human factors and ergonomics society annual meeting* 54, 25 (2010), 2130–2134.

[38] Jan L Plass, Paul A O'Keefe, Bruce D Homer, Jennifer Case, Elizabeth O Hayward, Murphy Stein, and Ken Perlin. 2013. The impact of individual, competitive, and collaborative mathematics game play on learning, performance, and motivation. *Journal of educational psychology* 105, 4 (2013), 1050.

[39] David V Pynadath, Ning Wang, and Sreekar Kamireddy. 2019. A Markovian Method for Predicting Trust Behavior in Human-Agent Interaction. In *Proceedings of the 7th International Conference on Human-Agent Interaction*. 171–178.

[40] SM Mizanoor Rahman, Yue Wang, Ian D Walker, Laine Mears, Richard Pak, and Sekou Remy. 2016. Trust-based compliant robot-human handovers of payloads in collaborative assembly in flexible manufacturing. In *2016 IEEE International Conference on Automation Science and Engineering (CASE)*. IEEE, 355–360.

[41] Paul Robinette, Michael Novitzky, Caileigh Fitzgerald, Michael R Benjamin, and Henrik Schmidt. 2019. Exploring human-robot trust during teaming in a real-world testbed. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 592–593.

[42] Julian Rode. 2010. Truth and trust in communication: Experiments on the effect of a competitive context. *Games and Economic Behavior* 68, 1 (2010), 325–338.

[43] Julian B Rotter. 1971. Generalized expectancies for interpersonal trust. *American psychologist* 26, 5 (1971), 443.

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

[45] 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.

[46] Nathan E Sanders and Chang S Nam. 2021. Applied quantitative models of trust in human-robot interaction. In *Trust in Human-Robot Interaction*. Elsevier, 449–476.

[47] Kristin Schaefer. 2013. *The perception and measurement of human-robot trust*. Ph. D. Dissertation. University of Central Florida.

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

[49] Isabel Schwaninger, Geraldine Fitzpatrick, and Astrid Weiss. 2019. Exploring trust in human-agent collaboration. In *Proceedings of 17th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET).

[50] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. 2019. "I don't believe you": Investigating the effects of robot trust violation and repair. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 57–65.

[51] Josh Siegel and Daniel Morris. 2020. Robotics, Automation, and the Future of Sports. *21st Century Sports: How Technologies Will Change Sports in the Digital*

Age (2020), 53–72.

[52] 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.

[53] 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.

[54] Rachel E. Stuck, Brittany E. Holthausen, and Bruce N. Walker. 2021. Chapter 8 - The role of risk in human-robot trust. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 179–194. https://doi.org/10.1016/B978-0-12-819472-0.00008-3

[55] 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.

[56] Anqi Xu and Gregory Dudek. 2015. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 221–228.

[57] Holly A Yanco, Jill L Drury, and Jean Scholtz. 2004. Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition. *Human–Computer Interaction* 19, 1-2 (2004), 117–149.

[58] Boling Yang, Xiangyu Xie, Golnaz Habibi, and Joshua R Smith. 2021. Competitive physical human-robot game play. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 242–246.

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