Multi-Contextual Analysis for Physiological Behaviour for Estimating Trust in Human-Robot Interaction

Abdullah S. Alzahrani[1,2] ORCID: 0009-0003-6036-1393 and Muneeb I. Ahmad[1] ORCID: 0000-0001-8111-9967

[1] Swansea University, Fabian Way, Crymlyn Burrows, Skewen, Swansea SA1 8EN
{2043528,m.i.ahmad}@swansea.ac.uk
https://www.swansea.ac.uk/compsci/
[2] Al-Baha University, 65779, Saudi Arabia
amisfer@bu.ed.sa

Abstract. Existing work on estimating user trust in robotic systems has primarily utilised datasets that monitored variations in physiological behaviours (PBs), evolving from one context of interaction. Consequently, in this paper, we created two datasets from two different human-robot interaction (HRI) contexts, namely competitive and collaborative, to explore trust dynamics comprehensively. The datasets consisted of participants' electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), skin temperature (SKT), blinking rate (BR), and blinking duration (BD) across multiple sessions of collaborative HRI during trust and distrust states. We investigated the differences in PBs between trust and distrust states and explored the potential of incremental transfer learning methods in predicting trust levels during HRI using the two datasets. The findings showed significant differences in HR between trust and distrust groups. It further showed that the Decision Tree classifier achieved the best accuracy of 89% in classifying trust, outperforming the previous work, while HR, BVP, and BR were the important features. Overall, the findings indicate the potential for applying incremental transfer learning to real-time datasets collected from different HRI contexts to estimate trust during HRI.

Keywords: Trust · Measurement · Physiological Behaviour · Human-Robot Interaction · Real-time.

1 Introduction

Trust in human-robot interaction (HRI) is a multifaceted psychological phenomenon involving beliefs about a robot's reliability and safety, shaped by interactions in uncertain situations [1]. Over the past decade, trust has gained significance due to the growing use of robots in fields like healthcare, economics, and industry, where it affects system acceptance and efficiency [45,15,13]. Measuring trust is challenging due to its dynamic nature and numerous influencing

factors [31]. Trust evolves through factors such as reliability, performance, error rates, and interaction context [49,5].

Traditional methods of measuring trust, including subjective methods (e.g., surveys, self-reports) and objective methods (e.g., performance metrics, behavioural analysis), often fail to capture trust's real-time, dynamic nature [31]. To address this, researchers have explored physiological measures, which enable real-time monitoring of dynamic trust [11]. Physiological behaviours (PBs) such as electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), blinking rate (BR), and blinking duration (BD) provide direct indicators of emotional and cognitive states during HRI [11,8].

In HRI, researchers have identified physiological indicators like EDA, HR, and SKT as predictors of trust using machine learning techniques [7,23,11]. However, traditional research often relies on data from a single experimental context, failing to capture the complexity of human trust across diverse environments and interactions. Additionally, most datasets are gathered in simulated settings that do not reflect the complex dynamics of trust and distrust. To address these gaps, this paper presents a dataset collected from multiple experimental contexts. Our experiment builds on our previous work [11] but alters the context. While their study gathered physiological signals in a competitive HRI setting, we collected a second dataset using the same signals in a collaborative HRI context.

This approach provided two datasets from different experimental contexts, allowing us to leverage knowledge from one context to enhance predictions in another, thereby aiming to improve the generalisability of our physiological findings across varied interaction scenarios. Using both datasets is essential for a comprehensive understanding and accurate trust prediction in HRI. By integrating data from competitive and collaborative contexts, we capture a wider range of interaction dynamics and physiological responses. Insights from one context can inform and improve predictions in the other, thereby enhancing the reliability of our trust classification models.

We applied multiple incremental transfer learning techniques to enhance model accuracy. This approach transfers pre-learned knowledge from one context (source domain) to related tasks (target domain), using multiple rounds with a negative transfer avoidance algorithm [52]. Although emerging in trust research within HRI, incremental transfer learning shows promise due to the diverse interaction dynamics in our datasets [19]. By leveraging this technique, we aim to improve model predictions across varied settings. Moreover, this study is distinct in utilising datasets from repeated competitive HRI and collecting an additional dataset in repeated collaborative contexts.

Considering these aspects, this paper investigates the following research questions (RQs):

- RQ1: Do PBs vary between trust and distrust cases during HRI in a repeated collaborative setting?
- RQ2: How does the interaction context (competitive and collaborative) affect PBs during trust and distrust states in repeated interactions?

– RQ3: How effective are incremental transfer learning techniques in improving the prediction accuracy of trust levels across different HRI contexts?

To address the RQs, we used the datasets from our previous work [11], which estimated human trust in robots by examining differences in PBs (EDA, BVP, HR, SKT, BR, BD) between trust and distrust during repeated competitive HRI. Their study found significant differences in HR and SKT between trust and distrust groups and achieved 68.6% accuracy using a Random Forest classifier. The second dataset, collected for this research, involved participants playing a collaborative game to decide whether to trust or distrust the NAO robot. Our approach stands out by using two datasets from distinct settings and applying incremental transfer learning to classify trust and distrust.

The contributions of this study are novel and valuable for future research in the field and are as follows:

– We show that HR vary between trust and distrust during collaborative HRI, providing insights into the dynamics of human trust in robots.
– We showed that there is a significant effect of interaction context (competitive and collaborative) on PBs during trust and distrust states, emphasising the need to consider contextual variation in trust modelling.
– We showed that incremental transfer learning techniques improve the predictive accuracy of trust models when combining data from different contexts (competitive and collaborative) of HRI.
– To further support and enable ongoing research in this field, we provide access to the study materials and the evolving dataset. These resources are made available to the academic community and can be accessed here.

The rest of this paper has been structured as follows: Section 2 provides an overview and discusses relevant literature. Section 3 provides a comprehensive description of the study. Sections 4 and 5 present the results and their discussion. Finally, section 6 concludes the paper.

2 Background & Related Work

2.1 Trust- Theoretical Understanding

The concept of trust varies across contexts and fields [34] and has been extensively studied in HRI [50]. However, a universal definition of trust is yet to be established [24]. Abbass et al. [1] defined trust as a multifaceted psychological phenomenon consisting of beliefs and expectations regarding trustworthiness, while Ajenaghughrure et al. [8] described it as a complex cognitive process operating subconsciously. These definitions highlight trust's dynamic nature and demonstrate the link between trust and human PBs, supporting the use of physiological signals for real-time trust assessment in robots.

Human PBs offer insights into emotional and cognitive states, making them valuable for real-time trust assessment. PBs like EDA, BVP, HR, SKT, BR,

and BD directly indicate physiological responses during robot interactions [43]. These can be measured via various sensors: EEG for neurological activity, HR and BVP for cardiac activity, EDA and SKT for skin responses, and BR and BD for ocular activity [6]. Studies have shown that specific PBs correlate with trust and distrust states, with increased HR and EDA often indicating distrust due to heightened arousal and stress [7,23,11].

2.2 Measuring Trust

Trust in HRI has been evaluated through subjective, objective, and physiological behaviour methodologies [32]. Subjective reporting was the most common approach, involving questionnaires and interviews [56,48,40]. Objective methods analyse human behaviour during interaction such as decisions and physical movement [42,53,3]. Trust can be accurately evaluated using PBs, which offer real-time, objective indicators of a person's emotional and cognitive states during interactions with robots [33].

Khalid et al. [30] explored psycho-physiological correlates of trust in human-robot-human interactions using heart rate variability (HRV), facial expressions, and voice features. Akash et al. [9] explored EEG and GSR for real-time trust sensors in human-machine interaction, with the general model achieving 70% accuracy and the customised model performing better at 78.58%. Lu and Sarter [38] studied eye movement as an indicator of trust in automation during a target identification task. Ajenaghughrure et al. [7] developed a model to evaluate human trust in conversational interfaces using physiological signals, achieving 77.8% accuracy. Hald et al. [23] designed a computer vision-based system to assess human-robot trust in real-time during close-proximity HRI, finding no significant differences in GSR between test conditions. Gupta et al. [22] investigated human trust in a virtual assistant using physiological sensing, finding HRV to be a reliable indicator of trust. Xu et al. [54] developed an EEG-based model to recognise trust in HRI, achieving 62% accuracy.

In summary, the current study on measuring human-robot trust through PBs and using machine-learning techniques for trust prediction presents both progress and limitations. Reported accuracy rates are typically high, often utilising direct supervised learning classification methods. Additionally, relying on data from a single dataset collected under a simulated environment fails to capture the full complexity and variability of real-world HRIs.

Our previous work [11] addressed the limitation of integrating multiple physiological behaviours in competitive HRI. Building on their work, we utilised their dataset and created a second dataset by modifying their experimental task for a collaborative context. We employed incremental transfer learning to combine insights from both datasets, enhancing the model's adaptability and significantly improving trust prediction accuracy in various HRI scenarios.

3 Research Method

This study aimed to measure human trust in robots through PBs in different HRI contexts using two distinct datasets. Firstly, we employed an existing dataset from our previous work [11], which focused on human trust and distrust

behaviours in a repeated competitive HRI setting. In that study, participants played a competitive card game against the NAO robot. During the game, participants had to decide whether to trust or distrust the robot's truthfulness (actions based on its statements). They measured participants' PBs, such as EDA, BVP, HR, SKT, BR, and BD, during each decision-making period.

Building on the previous work, this study introduced a complementary approach by shifting the context to a collaborative HRI setting, which involves participants working alongside the NAO robot in a cooperative task. In this new dataset, we followed a similar experimental procedure and measured the same PBs (EDA, BVP, HR, SKT, BR, and BD) during each decision-making moment to maintain consistency and comparability with the previous study. However, the key difference lies in the collaborative nature of the task, where participants and the robot worked toward a shared goal, fostering a cooperative dynamic rather than competition. This change in context allowed us to investigate how trust and distrust manifest differently across competitive and collaborative settings.

3.1 Incremental Transfer Learning for Trust Classification

To enhance predictive accuracy across the different HRI contexts, we employed an incremental transfer learning (ITL) approach, adapting the strategy from [17]. This involved initial model training on the source dataset (competitive HRI), followed by incremental adaptation using the target dataset (collaborative HRI). The ITL process iteratively uses models trained on source domain subsets to initialise training for corresponding target domain subsets, allowing knowledge retention while adjusting to the new context. Mechanisms to mitigate potential negative transfer, by assessing and addressing distributional differences between domains, were also incorporated. This methodology, while involving careful implementation, was chosen for its potential to improve model generalisability and performance by leveraging data from varied interaction scenarios, reflecting real-world complexities and fostering system adaptability across contexts. The specific application details of ITL and its impact on classifier performance are further elaborated in Section 4.2.

We investigate the following hypotheses:

H1 Human PBs, including EDA, BVP, HR, SKT, BR, and BD, will show significant differences between trust and distrust behaviours during interactions with a robotic agent in a collaborative setting.
H2 The interaction context (competitive and collaborative) significantly affects PBs during trust and distrust states.
H3 Incremental transfer learning will enhance the accuracy of models in predicting trust when combining datasets from collaborative and competitive HRI settings.

The Game - During each turn for Player 1 to decide, Player 1 discusses their decision-making with the NAO robot, seeking advice on whether to accept or reject Player 2's claim. The robot provides suggestions based on a pre-determined

strategy that can be accessed *here*. This strategy was applied consistently across all sessions to maintain consistency in the robot's advice. The robot's suggestions were presented in natural dialogue and followed the *Wizard of Oz (WOz)* methodology, where the participants were unaware of this control. If Player 1 follows the robot's advice, it is considered a trust case. If they ignore the advice, it is considered a distrust case, as shown in various studies [55,21,5].

In this study, we employed the **Bluff Game** that we developed earlier in our previous work [11]. The Bluff Game was used to create trust and distrust scenarios focused on the robot's truthfulness. In this version of the game, participants played against the NAO robot in a competitive setup. Each player (human and robot) received 15 cards from a deck of 52, consisting of four sets of aces, numbered cards 1-10, jacks, queens, and kings. The objective for both the human and the robot was to be the first to discard all their cards. During the game, players made statements about the cards they were discarding, and their opponents had to decide whether to trust or distrust the statement. If the statement was truthful and the opponent trusted it, the discard proceeded; if the statement was false and the opponent distrusted it, the cards were returned to the player who made the false claim. This gameplay created natural situations where trust and distrust were tested in a competitive setting.

In our study, we modified the game to create trust and distrust scenarios around the robot's trustworthiness in a collaborative setting, changing the nature of interaction from competition to cooperation. The updated game involves two players, Player 1 (human) and Player 2 (NAO robot), playing collaboratively as a team against an adversary agent. The deck remains the same with 52 cards, and each team starts with 15 cards. The goal remains to discard all cards first, but the dynamics now focus on trust between the human and the robot as teammates, where they must rely on each other's decisions and work together to outplay the adversary. The game is turn-based as follows:

1. Player 1's Turn (Participant and NAO Robot):
   - Player 1 selects a set of 2-4 cards to discard and declares their rank.
2. Player 2's Turn (Adversary Agent):
   - Player 2 (adversary agent) decides whether to accept or reject Player 1's claim.
3. Outcome Determination:
   - If Player 2 Accepts the Claim: The turn passes without revealing the cards.
   - If Player 2 Rejects the Claim: The cards are revealed.
     - If Player 1 was truthful, Player 2 must take the discarded cards.
     - If Player 1 was not truthful, Player 1 must take the cards back.

The game continues in turns until one team discards all their cards. The interactive interface provides play and decision buttons, enabling smooth interaction between the players and the game. The card list updates dynamically after each turn.

During each turn for Player 1 to decide, Player 1 discusses their decision-making with the NAO robot, seeking advice on whether to accept or reject

Player 2's claim. The robot provides suggestions based on a pre-determined strategy that can be accessed *here*. This strategy was applied consistently across all sessions to maintain consistency in the robot's advice. The robot's suggestions were presented in natural dialogue and followed the *Wizard of Oz (WOz)* methodology, where the participants were unaware of this control. If Player 1 follows the robot's advice, it is considered a trust case. If they ignore the advice, it is considered a distrust case, as shown in various studies [55,21,5].

Interaction Scenarios - The NAO robot was programmed to interact verbally with participants using the (*WOz*) technique, which allowed for operator control to ensure consistent behaviour across all sessions. We divided the interaction into three phases: introduction, gameplay, and closing. During the introduction phase of the first session, the NAO robot welcomed participants and instructed them to prepare for the game. This process was repeated in the second, third, and fourth sessions, with minor variations in the robot's speech content. In the gameplay phase, the robot verbally advised participants on whether to accept or reject the claims made by the other player and responded to the participants' decisions on whether to trust or distrust its advice. At the end of each session, the robot either congratulated or encouraged participants based on the outcome of the game. After the final session, the robot said goodbye and expressed hope to work with the participants again, indicating the end of the experiment. A detailed description of the interaction scenario can be found here for reusability purposes.

3.2 Participants

In this study, we employed a similar approach to recruit and classify participants in alignment with our previous study [11]. We initially enrolled 45 participants, but due to some data collection issues, the study countered only 41 participants. Participants were between the ages of 18 and 40 years old, with an average age of 30.45 years and a standard deviation of 4.14. Out of the 41 participants, 20 were female, 20 were male, and one chose not to disclose their gender. The participants were recruited via the university's electronic mailing system and flyers on the campus. They were able to book their study slots online using a scheduling platform called *Calendly*. The study protocol was approved by the university ethics board (reference number: 2202370516013).

3.3 Setup, Materials and Procedure

The study was conducted in two separate rooms, as illustrated in Figure 1. In Room 1, a laptop was placed on a table for participants to play the game, with the NAO robot [46] positioned beside them. Participants wore Pupil Invisible Eye Tracking Glasses [35] and the Empatica E4 wristband [27] to record PBs while seated in front of the robot. They also used a tablet to complete

demographic information. This room was designed to maintain consistent environmental conditions during the study, including steady room temperature and consistent lighting. This was done to minimise potential influences on physiological measures, specifically BD, BR, and SKT. In Room 2, the experimenter monitored the interaction and remotely controlled the robot from a laptop.

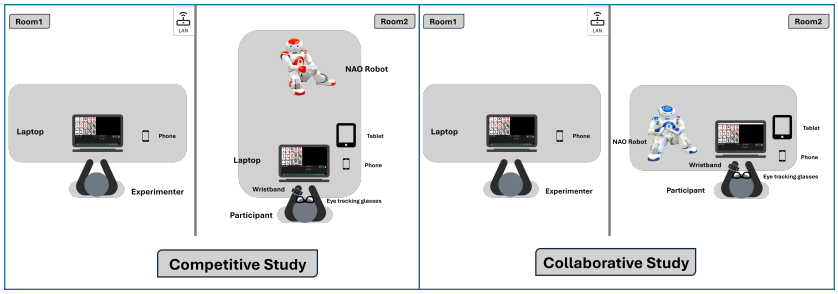

Fig. 1: Experiment Setup. Competitive setting in our previous study and setup in this study.

The procedure for the study, including participant instructions, game demonstration, and data collection process, can be seen in Figure 2. This figure outlines the key steps, from participant consent and demographic questionnaire completion to the game sessions and final debrief. Participants interacted with NAO robot across four sessions, with 5-mins breaks in between, and their PBs were recorded for analysis.

Fig. 2: Study Procedures

3.4 Measurements

In this study, we used the exact measurements and methods as in our previous study [11] to maintain consistency and build upon previous research. This approach ensures comparability of results and leverages the solid empirical evidence for trust measurement presented in the literature. We collected PBs, including EDA, BVP, HR, SKT, BR, and BD in real-time during decision periods, from when the robot played cards until the player made a decision. The choice of physiological signals prioritises participant comfort and non-intrusiveness. Wearable devices, specifically the Empatica E4 Wristband and the Pupil Invisible Eye

Tracking Glasses, capture chosen signals and have strong empirical evidence for trust measurement presented in the literature [11,16,57]. We also collected data on participants' in-game decisions, including their choices to trust or distrust the robot and each decision's start and end time. This information enabled us to assess the participants' PB responses during their decision.

In addition to physiological data, subjective trust ratings were collected to complement the objective measures. Following each interaction session in both the competitive (Study 1) and collaborative (Study 2) HRI contexts, participants completed the Trust Perception Scale-HRI (TPS-HRI) [48]. This scale was selected for its established validity in assessing human perceptions of trust in robotic agents. The data from the TPS-HRI were used to provide a comparative self-reported measure of trust, allowing for an examination of the convergence between subjective feelings and physiological responses.

3.5 Data Analysis

Preprocessing In order to create a dataset to assess human trust during HRI in real time using BPs, we performed the following steps:

1. **Decision outcomes logging:** 997 Trust and 306 distrust decisions made by participants were logged and coded as binary variables (0 for distrust, 1 for trust) for subsequent statistical analyses.
2. **Decision period logging:** Each participant's decision start and end time was logged to extract physiological data in the given interval. The gameplay log was maintained to extract the decision's start and end times. The start decision time is when the robot plays cards, and the end time is when the player presses one of the decision buttons.
3. **Noise and Artifact Removal:** The physiological data underwent an essential preprocessing step where we applied a Butterworth low-pass filter to remove noise and artifacts [41,14].
4. **Segmentation:** The physiological data was recorded with timestamps to mark the start and end of the session, allowing us to align it with the decision periods in the game. Subsequently, we segmented the data into four rounds for each participant.
5. **Feature extraction:** Based on each participant's decision start and the end time logged during the game, the physiological samples were aggregated by computing the average value of each PB. The raw physiological data, recorded at millisecond intervals, was averaged per second to make it more interpretable and reduce noise from rapid changes. This smoothing process allowed us to focus on meaningful, longer-term changes in PBs relevant to decision-making, enabling clearer comparisons during trust and distrust periods.

To ensure temporal coherence across the various physiological data streams, meticulous attention was paid to synchronisation. All data recording devices were synchronised at the commencement of each experimental session, and data alignment was further verified post-collection using timestamp analysis to ensure accurate correlation of physiological events.

Dataset Generation: To generate the dataset for the analysis and classification task in the study, we followed these steps:

1. First, we computed the value for each PBs during trust and distrust stated in all the sessions (1, 2, 3 and 4).
2. Next, in each session , we averaged the PBS for each participant in trust and distrust states. The normalisation was necessary due to variations in the number of trust and distrust decisions across participants in all four game sessions.
3. Later, all this resulted in a dataset containing 42 average values for each PB measure corresponding to trust and distrust decisions in each session.
4. Lastly, to form the dataset for all sessions, we merged the data of all the sessions into one.

By following these steps, we successfully generated a dataset suitable for analysing trust and distrust in HRI using PBs. The dataset alongside codes can be accessed here. In the given link, the file named as "Dataset1" represents the dataset of competitive study, while, the file named as "Dataset2" represents the dataset of collaborative study.

4 Results

To minimise individual variability, we collected data from participants during their rest period, which lasted 30 seconds at the beginning of each session. A repeated measures ANOVA was conducted to compare the mean values of all PBs between the baseline, trust, and distrust conditions across sessions in both datasets. The results showed a significant difference ($p < .05$) in PBs between the baseline and both trust and distrust conditions in both datasets. However, BVP did not show a significant difference, with p-values of .20 and .81 in the trust condition and .09 and .20 in the distrust condition.

To test **H1**, we performed a repeated-measures ANOVA to determine whether significant differences existed in the physiological measures (including EDA, BVP, HR, SKT, BR, and BD) depending on the decision (trust or distrust) and interactive session (session 1, session 2, session 3, or session 4).

We observed a significant effect of the decision on HR ($F(1, 71) = 15.346$**,** $p < .001$**,** $\eta_p^2 = .178$**.**) score. However, we did not observe a significant effect of the decision on EDA, BVP, SKT, BR, and BD.

No significant interaction effects between session and decision (session * decision) were found for any of the physiological measures, including HR, EDA, BVP, SKT, BR, and BD. This suggests that the influence of the decision (trust or distrust) on physiological responses did not vary across the four sessions.

We observed a significant main effect of session on **SKT** levels, $F(3, 69) = 22.599$, $p < .001$, $\eta^2 = .496$, indicating that skin temperature varied significantly across sessions.

A post-hoc Bonferroni test revealed that **SKT** was significantly higher in Session 1 compared to the other sessions: **Session 1 vs. Session 2**: $p < .001$,

with Session 1 showing a higher mean by 1.089 units. **Session 1 vs. Session 3**: $p < .001$, with Session 1 showing a higher mean by 1.130 units. **Session 1 vs. Session 4**: $p < .001$, with Session 1 showing a higher mean by 1.135 units.

However, no significant differences were found between Sessions 2, 3, and 4 ($p = 1.000$), suggesting stable SKT levels across these later sessions compared to Session 1. This pattern highlights a session-specific effect on SKT, particularly with Session 1 exhibiting consistently higher values.

The means and standard deviations for the physiological measures during trust and distrust across all sessions are presented in **Table 3**.

To test **H2**, we conducted a repeated-measures ANOVA to examine the effect of interaction setting (collaborative vs. competitive) on PBs (EDA, BVP, HR, SKT, BR, and BD). The results showed significant effects of Setting on EDA ($F(1, 155) = 5.071$, $p = 0.026$, $\eta^2 = 0.032$), BVP ($F(1, 155) = 6.282$, $p = 0.013$, $\eta^2 = 0.039$), HR ($F(1, 155) = 13.249$, $p < 0.001$, $\eta^2 = 0.079$), BR ($F(1, 155) = 192.188$, $p < 0.001$, $\eta^2 = 0.554$) and BD ($F(1, 155) = 205.616$, $p < 0.001$, $\eta^2 = 0.570$). However, we did not observe a significant effect of the setting on SKT.

Feature	N (Comp)	Trust (Comp)		Distrust (Comp)		N (Collab)	Trust (Collab)		Distrust (Collab)	
		M	SD	M	SD		M	SD	M	SD
EDA	43	0.86	2.09	0.85	2.02	41	0.38	0.44	0.36	0.40
BVP	43	0.13	0.99	0.14	0.87	41	0.01	0.14	0.00	0.10
HR	43	104.60	17.66	92.99	36.64	41	101.97	15.15	111.89	17.45
SKT	43	28.25	1.31	25.21	8.75	41	26.83	1.38	26.84	1.38
BR	43	1.55	2.89	1.46	3.37	41	8.67	4.95	8.96	5.58
BD	43	189.32	117.49	180.59	166.94	41	373.20	98.22	368.24	108.69

Table 1: Mean (M) and Standard Deviation (SD) for physiological features under trust and distrust conditions for competitive (comp) and collaborative (collab) studies.

		Session 1				Session 2				Session 3				Session 4			
		Trust		Distrust		Trust		Distrust		Trust		Distrust		Trust		Distrust	
Feature (Unit)	N	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
EDA (μS)	43	0.69	1.13	0.87	1.21	0.93	2.46	0.93	2.35	0.88	2.22	0.76	2.09	0.91	2.33	0.81	2.29
BVP (μV)	43	0.31	1.32	0.37	1.58	-0.01	0.24	0.00	0.16	-0.10	0.66	0.15	0.68	0.34	1.28	0.03	0.25
HR (bpm)	43	107.78	20.76	94.59	31.09	103.97	17.66	98.17	34.61	104.23	14.63	90.12	39.98	102.65	24.85	89.10	40.62
SKT ($^\circ C$)	43	27.80	1.22	25.39	7.21	28.20	1.30	26.22	7.36	28.49	1.27	24.66	10.12	28.54	1.36	24.58	10.09
BR (count)	43	1.56	1.62	1.11	1.26	1.54	3.62	1.80	5.38	1.39	3.63	2.00	1.82	1.75	3.79	1.74	3.49
BD (s)	43	177.21	149.85	181.86	164.68	197.11	114.75	151.66	145.33	191.83	99.70	183.96	180.80	191.15	102.08	204.91	176.27

Table 2: Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session in the Competitive setting.

Subjective Trust Ratings Analysis of the Trust Perception Scale-HRI (TPS-HRI) scores provided insights into participants' self-reported trust levels. In Study 1 (competitive context), there was a statistically significant increase in TPS-HRI scores across the four interaction sessions ($F(3, 126) = 3.47$, $p = .025$). The mean scores increased from Session 1 ($M = 0.782$, $SD = 0.141$) to Session

		Session 1				Session 2				Session 3				Session 4			
		Trust		Distrust		Trust		Distrust		Trust		Distrust		Trust		Distrust	
Feature (Unit)	N	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
EDA (μS)	41	0.37	0.45	0.28	0.26	0.39	0.44	0.40	0.43	0.37	0.43	0.38	0.43	0.37	0.46	0.40	0.47
BVP (μV)	41	0.03	0.11	0.01	0.08	0.02	0.07	0.00	0.02	-0.01	0.24	-0.02	0.19	0.01	0.07	0.02	0.07
HR (bpm)	41	103.35	17.42	112.94	19.21	103.50	18.53	112.88	16.52	101.23	11.17	110.77	14.53	99.79	12.40	110.70	19.51
SKT ($^{\circ}C$)	41	25.95	1.31	26.00	1.33	27.07	1.36	27.12	1.34	27.13	1.23	27.18	1.32	27.15	1.29	27.20	1.16
BR (count)	41	9.22	5.87	9.13	5.51	8.63	4.39	9.06	5.53	8.38	5.09	8.53	5.71	8.42	4.45	9.07	5.82
BD (s)	41	372.72	108.11	362.61	124.86	387.71	67.19	377.27	85.25	365.04	114.61	358.96	124.03	367.33	98.24	375.07	96.00

Table 3: Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session in the Collaborative setting.

4 (M = 0.862, SD = 0.125). A Pearson's correlation analysis revealed a strong, positive association between TPS-HRI scores and physiologically-labelled trust instances (r(2) = .968, p = .033).

Similarly, in Study 2 (collaborative context), a significant increase in TPS-HRI scores was observed across sessions (F(3, 120) = 3.89, p = .010), with means increasing from Session 1 (M = 0.793, SD = 0.132) to Session 4 (M = 0.872, SD = 0.118). A strong, positive correlation was found between TPS-HRI scores and physiologically-labelled trust instances (r(2) = .964, p = .036). These findings underscore the validity of our physiological trust annotations by demonstrating their alignment with participants' explicit trust evaluations.

To test **H3**, which aims to investigate the potential improvement of accuracy in predicting human trust through PBs by using incremental transfer learning, we started by using the traditional ML following a structured approach proposed by Ahmad et al. [6]. In this regard, we implemented seven different classifiers, including Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), AdaBoost (AB), Neural Network (NN) and Naive Bayes (NB). To evaluate the performance of these classifiers, we employed a 5-fold cross-validation technique. Our findings indicated that RF and LR yielded the highest accuracies of 69% and 65%, respectively. Nonetheless, it is noteworthy that the remaining classifiers also performed above chance level (see Table 4 for detailed results).

To further explore the accuracy findings, we have presented the classification report in Table 4 for all the classifiers, highlighting the F1 score for each class. The results indicate that RF achieved a comparatively higher accuracy when compared to other classifiers. Trust and distrust were predicted correctly at 66% and 69%, respectively, on this test data.

4.1 Comparison of Classification across Settings

It was observed that the RF classifier achieved the highest accuracy in both scenarios, with 69% in the collaborative setting and 68% in the competitive setting. Overall, as seen in table 5, classifiers in the collaborative setting generally demonstrated relatively higher accuracy compared to those in the competitive

setting. The collaborative setting also presented more balanced F1-scores between trust and distrust in contrast to the competitive setting. Furthermore, the collaborative setting exhibited more consistent performance across different classifiers, with all classifiers achieving above 60% accuracy. Conversely, the competitive setting displayed more variability in classifier performance, with lower accuracies and F1-scores.

Classifier	**Accuracy (%)**					**F1-scores**	
	Session 1	Session 2	Session 3	Session 4	All Sessions	Trust	Distrust
RF	67	60	59	50	**69**	**0.66**	**0.69**
LR	54	63	50	66	**65**	0.62	0.66
SVM	55	68	47	68	**64**	0.63	0.60
DT	64	51	54	43	**64**	0.61	0.64
AB	62	53	56	46	**65**	0.61	0.62
NN	55	62	55	51	**63**	0.58	0.60
NB	42	60	60	62	**62**	0.62	0.61

Table 4: Classifier Accuracy's and F1-scores for Physiological Behaviours in Trust Classification in a Collaborative HRI.

Classifier	**Accuracy (%)**					**F1-scores**	
	Session 1	Session 2	Session 3	Session 4	All Sessions	Trust	Distrust
SVM	61	39	49	44	**61**	0.66	0.56
RF	76	43	57	56	**68**	**0.75**	**0.58**
LR	60	41	44	43	**50**	0.51	0.49
DT	61	43	41	43	**60**	0.65	0.51
AB	71	41	55	57	**55**	0.58	0.52

Table 5: Classifier Accuracies and F1-scores for Physiological Behaviours in Trust Classification in a Competitive HRI [11].

4.2 Feature importance in the collaborative HRI

We investigated which PBs in our dataset were predictive of either trust or distrust by analysing the SHAP (SHapley Additive exPlanations) values. As the RF classifier displayed superior performance in predicting trust or distrust, we only present the feature importance of trust and distrust, respectively, in this classifier as shown in Figure 3.

Trust Classification Performance and Incremental Transfer Learning Application The classification models were evaluated on their ability to distinguish between trust and distrust states. The application of incremental transfer learning (ITL), as introduced in Section 3.1, played a crucial role in enhancing performance.

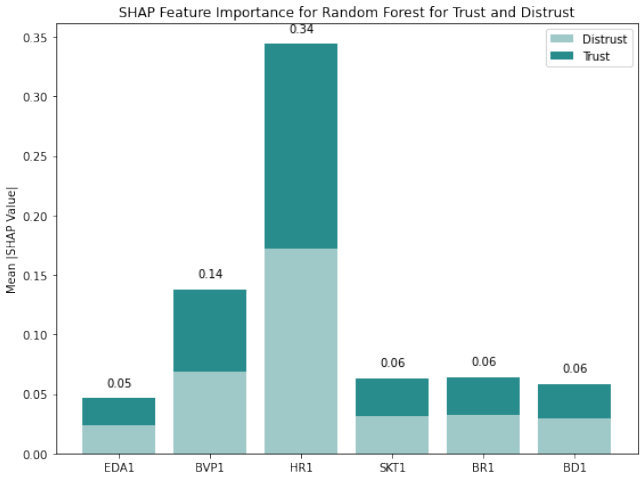

Fig. 3: Feature importance for the RF classifier based on SHAP mean value.

Specifically, the ITL strategy involved adapting models from the competitive (source) to the collaborative (target) HRI context. To simulate the sequential arrival of data as might occur in real-world scenarios, both the source and target datasets were partitioned into smaller, equally-sized subsets. In each iteration of the ITL process, a model trained on a data subset from the source domain was utilised to initialise the training of a new model on the corresponding subset from the target domain.

A key component of this approach was the active mitigation of potential negative transfer. This was achieved by employing Maximum Mean Discrepancy (MMD) to quantify the distributional divergence between the source and target domain subsets at each incremental step. If the calculated MMD value surpassed a predetermined threshold of 0.1, a sample reweighting strategy was implemented. This strategy assigned higher weights to source samples exhibiting greater similarity to the target distribution, with similarity assessed using a Radial Basis Function (RBF) kernel, thereby reducing the influence of less relevant source data.

Among the various classifiers evaluated, the Decision Tree classifier, when augmented with this ITL methodology, yielded the highest classification accuracy of 89%. This outcome represents a significant improvement over baseline models that did not incorporate transfer learning and also compares favourably with prior research typically limited to single-context datasets. The successful implementation of ITL thus highlights its considerable potential for developing more robust and adaptable trust estimation frameworks within HRI. (see Figure 4).

Subsequent analysis of feature importance for this best-performing model (Decision Tree with ITL) consistently identified HR, BVP, and BR as the most salient physiological indicators for differentiating trust and distrust states. A

preliminary interpretation suggests that HR variations likely reflect changes in emotional arousal and cognitive effort associated with trust decisions. BVP, an indicator of peripheral blood flow, is sensitive to autonomic nervous system responses often triggered by uncertainty or stress inherent in trust-related assessments. Furthermore, BR has been linked to attentional focus and cognitive load, potentially reflecting the decision-making processes involved when participants evaluate the robotś trustworthiness. These interpretations are further explored and substantiated with existing literature in the Discussion section.

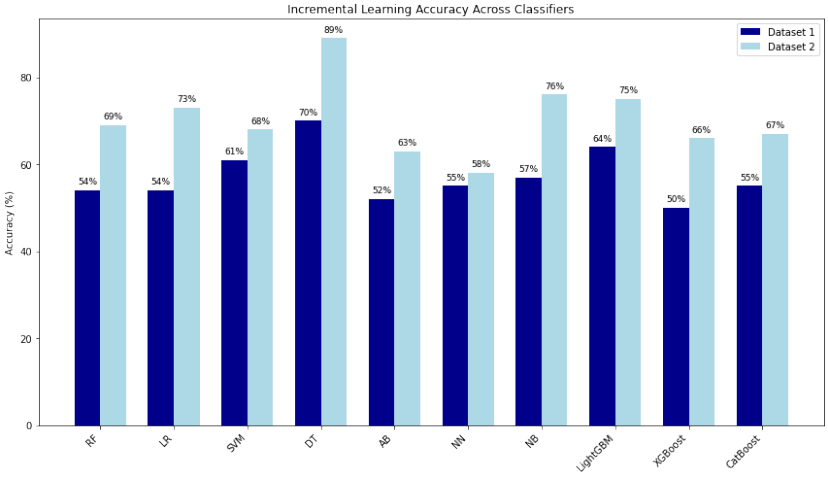

Fig. 4: Classification Accuracies Using Incremental Transfer Learning

5 Discussion

A key consideration in this research is the generalisability of the findings. While the experiments used a specific humanoid robot (NAO), our focus was on human physiological behaviours as indicators of internal states like trust. These behaviours possess a degree of universality, potentially less dependent on the robotic agent's specific morphology or speech content [2]. Our trust estimation models rely exclusively on signals from the human participant, rather than on robot characteristics. The deliberate incorporation of two distinct HRI contexts—competitive and collaborative—aimed to enhance model generalisability by exposing them to varied interaction dynamics.

H1 suggested that significant PBs differences exist between trust and distrust states, specifically in EDA, BVP, HR, SKT, BR, and BD. The results reveal that only HR exhibited a significant difference, indicating a robust association between heart rate and trust states. This finding is consistent with prior studies identifying HR as a critical indicator of assessing human trust in HRI contexts [8,30,44,22]. The outcome aligns with psychological theories suggesting

that trust-related emotional and cognitive processes significantly affect cardiovascular responses [10]. However, the absence of significant differences in EDA, BVP, SKT, BR, and BD suggests that these indicators may be less sensitive to trust-related emotional and cognitive states in this HRI setting, or these states may not have been strongly evoked by the experimental conditions. For instance, EDA has been shown to be a significant indicator of emotional arousal and cognitive effort in more intense or stress-inducing tasks, such as those involving time pressure or higher stakes [14]. Studies like Khawaji et al. [33] demonstrated significant changes in EDA when participants experienced high cognitive load during trust and distrust situations in a cooperative text-based task.

Trusting the robot in this context involves vulnerability, as participants need to rely on the robot's advice to achieve a successful outcome. This reliance can create heightened emotional arousal, as any perceived inconsistency or unreliability in the robot's behaviour could lead to increased stress and cognitive effort [28]. These factors contribute to a high HR in distrust situations, where participants might be more cautious and anxious about the robot's advice, reflecting the need for increased cognitive processing and emotional regulation [7]. Conversely, when participants trust the robot, their emotional state is likely more relaxed, resulting in a lower HR. Trust can help reduce the stress associated with the interaction, as participants feel more secure and confident in the robot's advice.

The absence of significant differences in EDA, BVP, BR, and BD suggests that these measures might be less sensitive to the variation of trust-related emotional states in the specific context of this experiment. These PBs may require more pronounced stressors or different types of cognitive tasks to show significant changes [14,22]. Additionally, the non-significant findings for these measures could also be due to individual differences in baseline physiological states or varying levels of responsiveness to the robot's actions among participants [51].

H2 hypothesised that there would be a significant effect of the settings (competitive vs. collaborative) on PBs. The findings confirmed this and showed a significant effect of the setting (competitive vs collaborative) on EDA, BVP, HR, BR, and BD. This shows that the context of interaction plays a crucial role in the factors influencing trust, as represented by PBs [12]. In competitive settings, the pressure to outperform the robot likely led to increased physiological arousal, as indicated by higher EDA, BVP, HR, BR, and BD. These conditions typically trigger stronger emotional and cognitive responses as participants aim to succeed against the robot, resulting in more noticeable physiological changes [20]. In contrast, working together in a collaborative setting may create a more cooperative and relaxed atmosphere, where both participants and robots work towards a common goal. This type of environment can help reduce stress and anxiety typically associated with competition, leading to lower physiological arousal [20]. The notable effects observed emphasize the importance of considering the interaction context when evaluating trust and PBs in HRI.

H3 indicated that using incremental transfer learning could enhance the accuracy of models in predicting trust by combining collaborative and competitive HRI datasets. Initially, we applied the direct ML method for classification. The

results with several classifiers demonstrate high accuracy in predicting trust. RF and LR classifiers yielded the highest accuracies, with RF achieving 69% and LR 65%, with HR, BVP, and BR features playing a crucial role in predicting human trust in robots during HRI. These features are closely linked to emotional arousal, cognitive effort, and rapid physiological changes that typically respond to trust-related decisions in gaming scenarios [37]. This finding is aligned with existing literature that demonstrates that PB features can predict trust in robots [26,30,7]. The best performance of RF, which uses multiple decision trees and majority voting, highlights its effectiveness in managing the complex, non-linear relationships inherent in physiological data. This robustness is attributed to the low correlation between the trees, enhancing the model's predictive power [39].

Notably, RF classifier achieved the highest accuracy in both collaborative and competitive settings, with accuracies of 69% and 68.6%, respectively. Classifiers generally performed slightly better in the collaborative setting, demonstrating higher accuracy and more balanced F1-scores between trust and distrust compared to the competitive setting. This consistency in collaborative scenarios suggests that PBs to trust are more stable and predictable when participants engage in cooperative tasks [47]. The higher performance in collaborative settings can be attributed to several factors. In collaborative tasks, participants and robots work towards a common goal, fostering a more relaxed and cooperative atmosphere. This environment likely reduces stress and anxiety, leading to more consistent and less variable physiological responses [18]. In contrast, competitive interactions introduce more variability in stress and anxiety levels, likely resulting in more diverse physiological responses, making it more challenging for classifiers to predict trust levels accurately [18].

The findings of the direct supervised learning classification support the potential of PBs in real-time trust assessment during HRI. The comparable performance of classifiers across both competitive and collaborative settings highlights the importance of considering interaction context when developing trust assessment models. This result, aligned with previous work, suggests that utilising multiple contexts in combination with incremental transfer learning can improve model robustness and adaptability [4]. Using this approach, we were able to establish more generalisable trust metrics across varied HRI scenarios.

H3 was accepted, demonstrating that incremental transfer learning effectively enhanced the classification accuracy when combining datasets from both collaborative and competitive HRI contexts. Specifically, the DT classifier achieved a 89% accuracy on the target dataset after integrating current data with our previous dataset [11]. While this level of accuracy might appear modest when compared to figures reported in some studies utilising different methodologies or sensor modalities, the principal contribution of this work extends beyond a singular accuracy metric. We have demonstrated the feasibility of using generalisable physiological signals, captured via non-intrusive wearable sensors, across two distinct HRI contexts (competitive and collaborative). This improvement can be attributed to the algorithm's ability to transfer relevant information while avoiding negative transfer [17]. By utilising diverse data from multiple in-

teraction contexts, we created a more comprehensive dataset that improved the models' generalisation and adaptability to new scenarios. These findings highlight the potential of incremental transfer learning in real-time trust assessment, supporting the development of adaptive robotic systems that can foster trust and enhance the effectiveness of HRIs.

The ability to estimate trust in real-time using physiological signals holds considerable promise for practical HRI applications. For instance, in collaborative manufacturing environments, cobots equipped with such trust-monitoring capabilities could dynamically adapt their behaviour based on the operatorś inferred trust state. This could involve modifications to operational speed, the predictability of movements, or even the proximity maintained with the human collaborator, thereby fostering safer and more efficient human-robot teamwork [25]. In assistive robotics, understanding the userś trust level could enable robots to adjust their level of autonomy or initiative, providing support that is better attuned to the userś comfort and confidence [29]. Similarly, in educational or therapeutic settings, a robot that can sense fluctuations in trust might adapt its interaction strategy to rebuild or maintain a positive rapport, enhancing the efficacy of the intervention [36]. The development of such adaptive systems, however, hinges on reliable and non-intrusive trust estimation, a goal towards which the present research contributes.

6 Conclusion and Future Work

This work demonstrates the potential of collecting physiological data in multiple contexts, showing that trust metrics based on PBs can be more robust and accurate by incorporating diverse interaction scenarios. This study analysed differences in PBs such as EDA, BVP, HR, SKT, BR, and BD between trust and distrust states in repeated collaborative HRI and explored the potential of combining PBs from different settings to enhance trust prediction accuracy using incremental transfer learning techniques. The findings confirmed that HR was a significant feature of trust. The findings also confirmed that the decision tree classifier achieved the highest accuracy of 89%. This technique successfully incorporated relevant information from varied datasets, enhancing model performance. Furthermore, the comparison between collaborative and competitive interactions revealed that the context of interaction significantly affects PBs associated with trust. This highlights the importance of considering the interaction context in trust assessment models.

It is important to acknowledge that while this study utilised two distinct HRI contexts, the specific characteristics of the chosen tasks and the participant sample may influence the broader generalisability of the findings.

In future work, we plan to explore how combining facial and speech features along with PBs in different contexts can improve trust prediction. We aim to use such approaches to develop adaptive robotic systems that can promote more meaningful and effective HRI.

References

1. Abbass, H.A., Scholz, J., Reid, D.J.: Foundations of Trusted Autonomy (Studies in Systems, Decision and Control). Springer Nature, Switzerland (2018)
2. Abubshait, A., Wiese, E.: You look human, but act like a machine: agent appearance and behavior modulate different aspects of human–robot interaction. Frontiers in psychology **8**, 1393 (2017)
3. Ahmad, M., Alzahrani, A., Robinson, S., Rahat, A.: Modelling human trust in robots during repeated interactions. In: Proceedings of the 11th International Conference on Human-Agent Interaction. pp. 281–290 (2023)
4. Ahmad, M., Alzahrani, A., Sunbul, A.: Detecting deception in natural environments using incremental transfer learning. In: INTERNATIONAL CONFERENCE ON MULTIMODAL INTERAC- TION (ICMI '24), November 4–8, 2024, San Jose, Costa Rica. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3678957.3685702 (2024)
5. Ahmad, M.I., Bernotat, J., Lohan, K., Eyssel, F.: Trust and cognitive load during human-robot interaction. arXiv preprint arXiv:1909.05160 (2019)
6. Ahmad, M.I., Keller, I., Robb, D.A., Lohan, K.S.: A framework to estimate cognitive load using physiological data. Personal and Ubiquitous Computing pp. 1–15 (2020)
7. Ajenaghughrure, I.B., Sousa, S.C., Kosunen, I.J., Lamas, D.: Predictive model to assess user trust: a psycho-physiological approach. In: Proceedings of the 10th Indian conference on human-computer interaction. pp. 1–10 (2019)
8. Ajenaghughrure, I.B., Sousa, S.D.C., Lamas, D.: Measuring trust with psychophysiological signals: a systematic mapping study of approaches used. Multimodal Technologies and Interaction **4**(3), 63 (2020)
9. Akash, K., Hu, W.L., Jain, N., Reid, T.: A classification model for sensing human trust in machines using eeg and gsr. ACM Transactions on Interactive Intelligent Systems (TiiS) **8**(4), 1–20 (2018)
10. Alexander, V., Blinder, C., Zak, P.J.: Why trust an algorithm? performance, cognition, and neurophysiology. Computers in Human Behavior **89**, 279–288 (2018)
11. Alzahrani, A., Ahmad, M.: Crucial clues: Investigating psychophysiological behaviors for measuring trust in human-robot interaction. In: Proceedings of the 25th International Conference on Multimodal Interaction. pp. 135–143 (2023)
12. Alzahrani, A., Robinson, S., Ahmad, M.: Exploring factors affecting user trust across different human-robot interaction settings and cultures. In: Proceedings of the 10th International Conference on Human-Agent Interaction. pp. 123–131 (2022)
13. Baratta, A., Cimino, A., Gnoni, M.G., Longo, F.: Human robot collaboration in industry 4.0: a literature review. Procedia Computer Science **217**, 1887–1895 (2023)
14. Boucsein, W.: Electrodermal activity. Springer Science & Business Media (2012)
15. Buxbaum, H., Sen, S., Kremer, L.: An investigation into the implication of human-robot collaboration in the health care sector. IFAC-PapersOnLine **52**(19), 217–222 (2019)
16. Chauhan, H., Pakbaz, A., Jang, Y., Jeong, I.: Analyzing trust dynamics in human–robot collaboration through psychophysiological responses in an immersive virtual construction environment. Journal of Computing in Civil Engineering **38**(4), 04024017 (2024)
17. Chui, K.T., Arya, V., Band, S.S., Alhalabi, M., Liu, R.W., Chi, H.R.: Facilitating innovation and knowledge transfer between homogeneous and heterogeneous

datasets: Generic incremental transfer learning approach and multidisciplinary studies. Journal of Innovation & Knowledge **8**(2), 100313 (2023)
18. De Francesco, L., Mazza, A., Sorrenti, M., Murino, V., Battegazzorre, E., Strada, F., Bottino, A.G., Dal Monte, O.: Cooperation and competition have same benefits but different costs. iScience (2024)
19. Diab, M., Demiris, Y.: A framework for trust-related knowledge transfer in human–robot interaction. Autonomous Agents and Multi-Agent Systems **38**(1), 24 (2024)
20. Gábana Arellano, D., Tokarchuk, L., Gunes, H.: Measuring affective, physiological and behavioural differences in solo, competitive and collaborative games. In: Intelligent Technologies for Interactive Entertainment: 8th International Conference, INTETAIN 2016, Utrecht, The Netherlands, June 28–30, 2016, Revised Selected Papers. pp. 184–193. Springer (2017)
21. Giorgi, I., Minutolo, A., Tirotto, F., Hagen, O., Esposito, M., Gianni, M., Palomino, M., Masala, G.L.: I am robot, your health adviser for older adults: do you trust my advice? International Journal of Social Robotics pp. 1–20 (2023)
22. Gupta, K., Hajika, R., Pai, Y.S., Duenser, A., Lochner, M., Billinghurst, M.: Measuring human trust in a virtual assistant using physiological sensing in virtual reality. In: 2020 IEEE Conference on virtual reality and 3D user interfaces (VR). pp. 756–765. IEEE (2020)
23. Hald, K., Rehmn, M., Moeslund, T.B.: Human-robot trust assessment using motion tracking & galvanic skin response. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 6282–6287. IEEE (2020)
24. Hancock, P.A., Kessler, T.T., Kaplan, A.D., Brill, J.C., Szalma, J.L.: Evolving trust in robots: specification through sequential and comparative meta-analyses. Human factors **63**(7), 1196–1229 (2021)
25. Hostettler, D., Mayer, S., Albert, J.L., Jenss, K.E., Hildebrand, C.: Real-time adaptive industrial robots: Improving safety and comfort in human-robot collaboration. arXiv preprint arXiv:2409.09429 (2024)
26. Hu, W.L., Akash, K., Jain, N., Reid, T.: Real-time sensing of trust in human-machine interactions. IFAC-PapersOnLine **49**(32), 48–53 (2016)
27. Inc., E.: Empatica: Wearable devices for health monitoring (2025), `https://www.empatica.com/`, accessed: 05-February-2025
28. Karakikes, M., Nathanael, D.: The effect of cognitive workload on decision authority assignment in human–robot collaboration. Cognition, Technology & Work **25**(1), 31–43 (2023)
29. Karim, R., Nanavati, A., Faulkner, T.A.K., Srinivasa, S.S.: Investigating the levels of autonomy for personalization in assistive robotics (2023)
30. Khalid, H.M., Shiung, L.W., Nooralishahi, P., Rasool, Z., Helander, M.G., Kiong, L.C., Ai-vyrn, C.: Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 60, pp. 697–701. SAGE Publications Sage CA: Los Angeles, CA (2016)
31. Khavas, Z.R.: A review on trust in human-robot interaction. arXiv preprint arXiv:2105.10045 (2021)
32. Khavas, Z.R., Ahmadzadeh, S.R., Robinette, P.: Modeling trust in human-robot interaction: A survey. In: Wagner, A.R., Feil-Seifer, D., Haring, K.S., Rossi, S., Williams, T., He, H., Sam Ge, S. (eds.) Social Robotics. pp. 529–541. Springer International Publishing, Cham (2020)
33. Khawaji, A., Zhou, J., Chen, F., Marcus, N.: Using galvanic skin response (gsr) to measure trust and cognitive load in the text-chat environment. In: Proceedings

of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. pp. 1989–1994 (2015)
34. Kok, B.C., Soh, H.: Trust in robots: Challenges and opportunities. Current Robotics Reports **1**(4), 297–309 (2020)
35. Labs, P.: Pupil labs invisible: Eye-tracking product (2025), `https://pupil-labs.com/products/invisible/`, accessed: 05-February-2025
36. Langer, A., Feingold-Polak, R., Mueller, O., Kellmeyer, P., Levy-Tzedek, S.: Trust in socially assistive robots: Considerations for use in rehabilitation. Neuroscience & Biobehavioral Reviews **104**, 231–239 (2019)
37. Lerner, J.S., Li, Y., Valdesolo, P., Kassam, K.S.: Emotion and decision making. Annual review of psychology **66**, 799–823 (2015)
38. Lu, Y., Sarter, N.: Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability. IEEE Transactions on Human-Machine Systems **49**(6), 560–568 (2019)
39. Mahesh, B.: Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet] **9**, 381–386 (2020)
40. Malle, B.F., Ullman, D.: A multidimensional conception and measure of human-robot trust. In: Trust in human-robot interaction, pp. 3–25. Elsevier (2021)
41. Mello, R.G., Oliveira, L.F., Nadal, J.: Digital butterworth filter for subtracting noise from low magnitude surface electromyogram. Computer methods and programs in biomedicine **87**(1), 28–35 (2007)
42. Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., Ju, W.: Behavioral measurement of trust in automation: the trust fall. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 60, pp. 1849–1853. SAGE Publications Sage CA: Los Angeles, CA (2016)
43. Moini, J., LoGalbo, A., Ahangari, R.: Foundations of the Mind, Brain, and Behavioral Relationships: Understanding Physiological Psychology. Elsevier (2023)
44. Nahavandi, S.: Trust in autonomous systems-itrust lab: Future directions for analysis of trust with autonomous systems. IEEE Systems, Man, and Cybernetics Magazine **5**(3), 52–59 (2019)
45. Renteria, A., Alvarez-de-los Mozos, E.: Human-robot collaboration as a new paradigm in circular economy for weee management. Procedia Manufacturing **38**, 375–382 (2019)
46. Robotics, A.: Nao robot official website (2025), `https://www.aldebaran.com/en/NAO`, accessed: 05-February-2025
47. Safryghin, A., Hebesberger, D.V., Wascher, C.A.: Individual behavioral and physiological responses during different experimental situations–consistency over time and effects of context. bioRxiv p. 477638 (2018)
48. Schaefer, K.E.: Measuring trust in human robot interactions: Development of the "trust perception scale-hri". In: Robust intelligence and trust in autonomous systems, pp. 191–218. Springer (2016)
49. Schaefer, K.E., Chen, J.Y., Szalma, J.L., Hancock, P.A.: A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human factors **58**(3), 377–400 (2016)
50. Schroepfer, P., Pradalier, C.: Why there is no definition of trust: A systems approach with a metamodel representation. In: 2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN). pp. 1245–1251. IEEE (2023)
51. Stemmler, G., Wacker, J.: Personality, emotion, and individual differences in physiological responses. Biological psychology **84**(3), 541–551 (2010)

52. Wang, Z., Dai, Z., Póczos, B., Carbonell, J.: Characterizing and avoiding negative transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11293–11302 (2019)
53. Wright, T.J., Horrey, W.J., Lesch, M.F., Rahman, M.M.: Drivers' trust in an autonomous system: Exploring a covert video-based measure of trust. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. vol. 60, pp. 1334–1338. SAGE Publications Sage CA: Los Angeles, CA (2016)
54. Xu, C., Zhang, C., Zhou, Y., Wang, Z., Lu, P., He, B.: Trust recognition in human-robot cooperation using eeg. arXiv preprint arXiv:2403.05225 (2024)
55. Xu, J., Howard, A.: How much do you trust your self-driving car? exploring human-robot trust in high-risk scenarios. In: 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 4273–4280. IEEE (2020)
56. Yagoda, R.E., Gillan, D.J.: You want me to trust a robot? the development of a human–robot interaction trust scale. International Journal of Social Robotics **4**, 235–248 (2012)
57. Yan, Y., Su, H., Jia, Y.: Measuring human comfort in human-robot collaboration via wearable sensing. IEEE Transactions on Cognitive and Developmental Systems (2024)