

Measuring Human's Trust in Robots in Real-time During Human-Robot Interaction



Swansea University
Prifysgol Abertawe

Abdullah Saad Alzahrani

Submitted in fulfilment of the requirements
for the Degree of Doctor of Philosophy

Department of Computer Science
Faculty of Science and Engineering
Swansea University

2025

To my father, who I dearly miss, and
my mother, who has been my rock.

"Behind every great achievement is a great belief"

- Ghazi Al Gosaibi

Abstract

This thesis presents a novel, holistic framework for understanding, measuring, and optimising human trust in robots, integrating cultural factors, mathematical modelling, physiological indicators, and behavioural analysis to establish foundational methodologies for trust-aware robotic systems. Through this comprehensive approach, we address the critical challenge of trust calibration in human-robot interaction (HRI) across diverse contexts. Trust is essential for effective HRI, impacting user acceptance, safety, and overall task performance in both collaborative and competitive settings. This thesis investigated a multi-faceted approach to understanding, modelling, and optimising human trust in robots across various HRI contexts. First, we explored cultural and contextual differences in trust, conducting cross-cultural studies in Saudi Arabia and the United Kingdom. Findings showed that trust factors such as controllability, usability, and risk perception vary significantly across cultures and HRI scenarios, highlighting the need for flexible, adaptive trust models that can accommodate these dynamics. Building on these cultural insights as a critical dimension of our holistic trust framework, we developed a mathematical model that emulates the layered framework of trust (initial, situational, and learned) to estimate trust in real-time. Experimental validation through repeated interactions demonstrated the model's ability to dynamically calibrate trust with both trust perception scores (TPS) and interaction sessions serving as significant predictors. This model showed promise for adaptive HRI systems capable of responding to evolving trust states. To further enhance our comprehensive trust measurement approach, this thesis explored physiological behaviours (PBs) as objective indicators. By using electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), skin temperature (SKT), eye blinking rate (BR), and blinking duration (BD), we showed that specific PBs (HR, SKT) vary between trust and distrust states and can effectively predict trust levels in real-time. Extending this approach, we compared PB data across competitive and collaborative contexts and employed incremental transfer learning to improve predictive accuracy.

across different interaction settings. Recognising the potential of less intrusive trust indicators, we also examined vocal and non-vocal cues—such as pitch, speech rate, facial expressions, and blend shapes—as complementary measures of trust. Results indicated that these cues can reliably assess current trust states in real-time and predict trust development in subsequent interactions, with trust-related behaviours evolving over time in repeated HRI sessions. Our comprehensive analysis demonstrated that integrating these expressive behaviours provides quantifiable measurements for capturing trust, establishing them as reliable metrics within real-time assessment frameworks. As the final component of our integrated trust framework, this thesis explored reinforcement learning (RL) for trust optimisation in simulated environments. Integrating our trust model into an RL framework, we demonstrated that dynamically calibrated trust can enhance task performance and reduce the risks of both under and over-reliance on robotic systems. Together, these multifaceted contributions advance a holistic understanding of trust measurement and calibration in HRI, encompassing cultural insights, mathematical modelling, physiological and expressive behaviour analysis, and adaptive control. This integrated approach establishes foundational methodologies for developing trust-aware robots capable of enhancing collaborative outcomes and fostering sustained user trust in real-world applications. The framework presented in this thesis represents a significant advancement in creating robotic systems that can dynamically adapt to human trust states across diverse contexts and interaction scenarios

Acknowledgements

All praise is due to Almighty Allah, the Most Gracious, the Most Merciful. His blessings, guidance, and grace have enabled me to undertake and complete this journey. None of this would have been possible without His infinite mercy and support.

I wish to express my heartfelt and deepest gratitude to my supervisor, Dr Muneeb Imtiaz Ahmad. His exceptional mentorship, tireless support, and unwavering dedication have been instrumental in the completion of this research. From countless hours of in-depth discussions in his office to late-night meetings ahead of paper submissions, his commitment to my progress has never wavered. His follow-ups, encouragement, and insightful advice have profoundly shaped both my academic and personal development, and for this, I am forever grateful.

I am also sincerely thankful to my co-supervisor, Professor Simon Robinson, for his thoughtful guidance and constructive feedback throughout my doctoral studies. His contributions have played a significant role in refining my thinking and improving the quality of my work.

I would like to extend my sincere thanks to my examiners, Dr Patrick Holthaus and Dr Sean Walton, for their time, effort, and valuable feedback during my viva examination. Their insights and thoughtful questions have contributed meaningfully to the final shape of this thesis, and I am grateful for their constructive engagement.

I gratefully acknowledge the generous sponsorship and financial support provided by the Saudi Government, represented by Al-Baha University. This opportunity has been a cornerstone of my academic and professional development, and I am truly thankful for the trust placed in me.

I am forever indebted to my mother, Atiqah Ibrahim, whose unconditional love, prayers, and unwavering support have been a constant source of strength and inspiration throughout my life.

To my beloved wife, Fatimah, and my children, Hanay and Hani, I express my deepest gratitude for their patience, love, and understanding. Their steadfast support and encouragement have sustained me during the most challenging moments of this journey.

I also extend my heartfelt thanks to my brothers and sisters for their continuous encouragement and belief in my abilities. Their support has been a pillar of strength throughout this endeavour.

I am sincerely grateful to my dear friends — Dr Abdulkareem Aodah, Ahmad Alharbi, Omer Aziz, Abdulrahman Khalaf, Ahmad Misfer, Tareq Taher, and Abdulrahman Kamal — whose friendship and encouragement have been a great source of comfort and motivation.

My thanks also go to all my colleagues and collaborators at Swansea University. Their insights, discussions, and companionship have significantly enriched my research experience and made this journey both productive and memorable.

Lastly, I extend my gratitude to everyone who has contributed to this work in any way, whether through direct involvement or emotional support. This journey has truly been a collective effort, and I am deeply grateful for the kindness and encouragement I have received along the way.

Declaration of Authorship

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree. However, the work in this thesis has undergone peer review and has been presented at major international venues. The publications noted below were prepared for submission during my candidature.

- Accepted Papers:

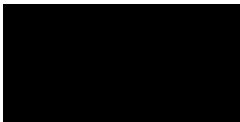
- (P1) **Abdullah Alzahrani**, Simon Robinson, and Muneeb Ahmad. 2022. *Exploring Factors Affecting User Trust Across Different Human-Robot Interaction Settings and Cultures*. In *Proceedings of the 10th International Conference on Human-Agent Interaction (HAI '22)*. Association for Computing Machinery, New York, NY, USA, 123–131. <https://doi.org/10.1145/3527188.3561920>
- (P2) **Abdullah Alzahrani** and Muneeb Ahmad. 2023. *Crucial Clues: Investigating Psychophysiological Behaviors for Measuring Trust in Human-Robot Interaction*. In *Proceedings of the 25th International Conference on Multimodal Interaction (ICMI '23)*. Association for Computing Machinery, New York, NY, USA, 135–143. <https://doi.org/10.1145/3577190.3614148>
- (P3) Muneeb Ahmad, **Abdullah Alzahrani**, Simon Robinson, and Alma Rahat. 2023. *Modelling Human Trust in Robots During Repeated Interactions*. In *Proceedings of the 11th International Conference on Human-Agent Interaction (HAI '23)*. Association for Computing Machinery, New York, NY, USA, 281–290. <https://doi.org/10.1145/3623809.3623892>
- (P4) **Abdullah Alzahrani** and Muneeb Ahmad. 2024. *An Estimation of Three-Layered Human's Trust in Robots*. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24 Companion)*. Association for Computing Machinery, New York, NY, USA, 144–146. <https://doi.org/10.1145/3640544.3645242>

- (P5) *Muneeb Imtiaz Ahmad, Abdullah Alzahrani, and Sunbul M. Ahmad. 2024. Detecting Deception in Natural Environments Using Incremental Transfer Learning. In Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24). Association for Computing Machinery, New York, NY, USA, 66–75. <https://doi.org/10.1145/3678957.3685702>*
- (P6) *Abdullah Alzahrani and Muneeb Ahmad. 2024. Real-Time Trust Measurement in Human-Robot Interaction: Insights from Physiological Behaviours. In Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24). Association for Computing Machinery, New York, NY, USA, 627–631. <https://doi.org/10.1145/3678957.3688620>*
- (P7) *Abdullah Alzahrani, Jauwairia Nasir, Elisabeth André, Ahmad J. Tayeb, and Muneeb Ahmad. 2025. What Do the Face and Voice Reveal? Investigating Trust Dynamics During Human-Robot Interaction. In Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25). Association for Computing Machinery, New York, NY, USA, 32–41.*
- (P8) *Abdullah Alzahrani, and Muneeb Ahmad. Optimising Human Trust in Robots: A Reinforcement Learning Approach. In Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25). Association for Computing Machinery, New York, NY, USA, 32–41.*
- Under review papers:

(P9) *Abdullah Alzahrani, Simon Robinson and Muneeb Ahmad. A Three-Layered Framework for Estimating Human Trust in Robots During Repeated Interactions. International Journal of Social Robotics. [Currently under-review after receiving 1 Minor and 2 Major Revisions in April 2024].*

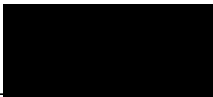
(P10) *Abdullah Alzahrani, and Muneeb Ahmad. Multi-Contextual Analysis for Physiological Behaviour for Estimating Trust in Human-Robot Interaction.*

DECLARATION - This work has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed  (candidate)

Date 09/05/2025

Statement 1 - This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed  (candidate)

Date 09/05/2025

Statement 2 - I hereby give consent for my thesis, if accepted, to be available for electronic sharing.

Signed  (candidate)

Date 09/05/2025

Statement 3 - The University's ethical procedures have been followed and, where appropriate, that ethical approval has been granted.

Signed  (candidate)

Date 09/05/2025

Contents

Abstract	iii
Acknowledgements	v
Declaration of Authorship	vii
1 Introduction	1
1.1 Motivation	2
1.2 Research Objective	5
1.3 Contributions	8
1.4 Thesis Outline	10
2 Review of Related Work	12
2.1 Theoretical Understanding of Trust	12
Cognitive Expectations: Trust as a Mental Construct	13
Physiological Manifestations: Trust as a Subconscious Re- sponse	14
Contextual Dynamics: Trust as Situational and Adaptive	14
2.2 Factors Affecting Human Trust in Robots	15
2.2.1 Trust transfer across settings	17
Trust in Human-robot Competition	18
2.2.2 Culture and trust	19
2.3 Measuring Human Trust in Robots	21
2.3.1 Mathematically modelling	21
2.3.2 Physiological Behaviour Measurement	25
Empirical Evaluations and Physiological Behaviours	26
Machine learning to estimate Trust	28
2.3.3 Trust And Vocal and Non-vocal cues	33
2.4 Trust in Repeated or Long-term interactions	35

2.4.1	Trust Framework	35
2.4.2	Empirical Studies on Trust in Repeated or Long-term Inter- actions	36
2.5	Reinforcement Learning For Optimising Trust in HRI	40
3	Factors Affecting Human Trust in Robots	43
3.1	Study Design	45
3.1.1	Ethics	46
3.1.2	Task	46
3.1.3	Participants	47
3.1.4	Procedure	49
3.1.5	Measurements	49
3.2	Results	50
3.2.1	Quantitative findings	50
	Study 1 (Saudi Arabia)	50
	Study 2 (United Kingdom)	53
	Results – comparing Studies 1 and 2	53
3.2.2	Qualitative findings	54
	Study 1 (KSA)	55
	Study 2 (United Kingdom)	58
	Results – comparing studies 1 and 2	59
3.3	Discussion	60
	Quantitative findings	60
	Qualitative findings	63
3.4	Conclusions, limitations and future work	64
4	Mathematical Trust Model	66
4.1	Initial Trust Model	69
4.2	Formative Evaluation of Approach	72
4.2.1	Ethics	72
4.2.2	System description	72
	The Game	73
	Interaction Scenarios	75
4.2.3	Participants	76
4.2.4	Setup and Materials	77
4.2.5	Procedure	78

4.2.6	Measurements	78
4.3	Results of Trust Model Formative Evaluation	79
4.3.1	Discussion and Implications of Formative Evaluation	84
4.4	Extended model	85
4.5	Summative Evaluation of the Model	88
4.5.1	Ethics	89
4.5.2	System description	90
	The Game	91
	Interaction Scenarios	94
4.5.3	Participants	96
4.5.4	Setup and Materials	97
4.5.5	Procedure	97
4.5.6	Measurements	99
4.6	Results of Summative Evaluation	99
4.6.1	H1: Predicting TMS with TPS and Session	100
4.6.2	H2: The Effect of Interactive Sessions on TPS and TMS	101
4.6.3	H3: Differences Across Trust Layers	102
4.6.4	Comparison of Initial and Refined Trust Models	103
4.7	Third Study	103
4.7.1	System Description	104
4.7.2	The Game: Matching Pairs	105
4.7.3	Participants	108
4.7.4	Setup and Materials	108
4.7.5	Procedure	109
4.7.6	Measurements	109
4.8	Results	110
4.8.1	H4: Predicting TMS with TPS and Session	110
4.8.2	H5: The Effect of Interactive Sessions on TPS and TMS	110
4.8.3	H6: Differences Across Trust Layers	111
4.9	Discussion	112
4.10	Conclusion	116
5	Human Physiological Behaviours as Indicators of Trust in Robots	118
5.1	First Study	120
5.1.1	Data Collection	121

	Behavioural Measures	121
5.1.2	Data Preparation	122
	Behavioural Data Processing	122
	Physiological Data Preprocessing	122
5.2	Results	124
	Overfitting Analysis	128
5.2.1	Feature importance for Trust and Distrust	131
5.2.2	Comparison of the models' performance	132
5.3	The Second Study	134
5.3.1	Data Collection	135
	Preprocessing	135
	Dataset Generation:	136
5.4	Results	137
5.4.1	Hypothesis 1 (H1) testing	137
	Main Effects of Decision:	137
	Interaction Effects:	137
	Main Effects of Session:	137
	Post-Hoc Analysis:	137
5.4.2	Hypothesis 2 (H2) testing	138
5.4.3	Hypothesis 3 (H3) testing	142
	Overfitting Analysis for Collaborative Setting	143
5.4.4	Comparison of Direct Supervised Learning Classification in Competitive vs. Collaborative Settings	147
5.4.5	Feature importance for Trust and Distrust	148
5.4.6	Incremental Transfer Learning Results	148
	Overfitting Analysis	150
	Comparison of the models performance	153
5.5	Discussion	154
5.5.1	Effect of Decision	154
5.5.2	Interaction Effect	155
5.5.3	Classification Accuracy	156
5.5.4	Setting Effect	156
5.5.5	Incremental Transfer Learning	157
5.6	Conclusion	157

6	Predicting Human Trust in Robot Using Vocal and Non-vocal Cues	159
6.1	User Study	161
6.1.1	Hypotheses	161
6.1.2	Procedure	161
6.1.3	Data Collection	161
	Facial Expression and Blendshape	162
	Voice Feature	164
6.1.4	Data Analysis	165
6.2	Results	166
6.2.1	Construct Validity with Questionnaire	166
6.2.2	Effect of Session & Decision on the Human Cues	167
6.2.3	Classification of Trust and Distrust Decisions	167
6.2.4	Addressing Overfitting in Classification Models	171
	Overfitting Analysis	171
	Regularisation Techniques	172
	Results of Regularisation	173
	Confusion Matrix Analysis	175
6.3	Discussion	176
6.3.1	Differences in Vocal and Non-Vocal Behaviours Between Trust and Distrust States (H1)	176
6.3.2	Effect of Interaction (Decision*Session) on Vocal and Non-Vocal Behaviours(H2)	177
6.3.3	Accuracy of Machine Learning Classifiers (H3)	178
6.3.4	Most Predictive Vocal and Non-Vocal Behaviours (H4)	178
6.4	Conclusion and Limitation	179
7	Reinforcement Learning	181
7.1	Methodology	182
7.1.1	Markov Decision Process (MDP) Framework	182
	State Space (S)	183
	Action Space (A)	183
	Transition Probabilities (P)	184
	Reward Function (R)	184
	Discount Factor (γ)	185
7.1.2	Simulation Environments	185

Frozen Lake Environment	185
Battleship Environment	186
7.1.3 Q-Learning Algorithm	186
Q-Table Structure	186
Action Selection	186
Q-Value Update Rule	186
7.1.4 Trust Update Mechanism	187
7.2 Results and Discussion	188
7.2.1 Performance Evaluation	188
7.2.2 Trust State Analysis	190
7.3 Conclusion and Future Work	192
8 Discussion & Conclusions	193
8.1 Open Questions	193
8.2 Conclusion	195
A Supplementary Material	197
Study 2: Game Strategy	208
Study 3: Game Strategy	218
Bibliography	223

List of Figures

1.1	Thesis Road-map	11
2.1	Hancock's three factors of trust model	16
3.1	The steps taken in the workshop (upper left to lower right).	47
3.2	Histograms and Q-Q plots for TPS	51
3.3	Histograms and Q-Q plots for TPS	52
3.4	Commonalities and differences among factors affecting human trust in robots across three scenarios in Studies 1 & 2.	56
4.1	Modelling the Three Layers of Trust.	69
4.2	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.	71
4.3	System architecture for the trust measurement platform. The sys- tem integrates multiple components: (1) a card game module that creates trust-based decision scenarios, (2) a semi-autonomous NAO robot that interacts with participants, (3) a trust computation mod- ule implementing our mathematical model, and (4) a data collec- tion system capturing interaction patterns. The architecture en- ables real-time trust measurement during HRI in a competitive game setting.	73
4.4	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).	77
4.5	Scatter plot and linear regression line showing a relationship be- tween the computed trust modelled score and the predicted trust modelled score based on the trust perception score and time.	82

4.6	Scatter plot illustrating the evolution of trust over four interactive sessions. Trust Perception Score (TPS) is shown in green, representing participants' subjective trust ratings, while Trust Modelled Score (TMS) is shown in orange, representing the trust computed using the mathematical model. The trend lines indicate a gradual increase in both TPS and TMS, suggesting that trust develops positively over time.	83
4.7	Illustration of the impact of Current Trust Levels $T(t)$ and Experiences $E(t)$ on the New Trust Level $T(t + \Delta t)$ for $\gamma = 0.25$, showing that a highly positive experience has a limited impact when current trust is low.	88
4.8	Overview of the trust-adaptive system. The system comprises two integrated modules: (1) an interactive Bluff Card Game designed to elicit varying levels of trust and distrust by placing participants in decision-making scenarios under conditions of uncertainty and risk, and (2) a semi-autonomous robotic partner that provides advice and engages dynamically with the participant during gameplay. The system captures behavioural, physiological, and decision-based responses to investigate the influence of robot advice accuracy, participant control, and perceived risk on trust formation and calibration. This setup enables real-time assessment and modelling of human trust dynamics in repeated HRI.	90
4.9	Experiment Setup. An experimenter controls the robot in one room (left), while the participant is playing the game with the assistance of the robot in another room (right).	98
4.10	A regression plot displaying the relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and session variables.	101
4.11	Scatter plot depicting the changes in the trust perception score (in Orange) and trust modelled score (in Blue) over time.	102
4.12	Comparison of regression lines for the initial (blue) and refined (green) trust models, illustrating the improved predictive capability of the refined model in estimating the TMS.	104
4.13	Experiment System	105

4.14	Experiment Setup. An experimenter controls the robot in one room (left) while the participant plays the game with the assistant of the NAO robot in another room (right).	109
4.15	A regression plot displaying the relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and session variables.	111
4.16	Scatter plot depicting the changes in the trust perception score (in Orange) and trust modelled score (in Blue) over time.	112
5.1	Empatica E4 Wristband (left) and Pupil Invisible Eye Tracking Glasses (right).	121
5.2	Learning curve for the Random Forest classifier showing training and cross-validation scores as a function of training set size. The minimal gap between training and validation scores demonstrates excellent generalisation capabilities with no overfitting.	129
5.3	Confusion matrix for the Random Forest classifier showing the distribution of true positives, false positives, true negatives, and false negatives.	130
5.4	ROC curve for the Random Forest classifier showing the trade-off between true positive rate and false positive rate at different classification thresholds. The AUC of 0.85 indicates excellent discriminative ability.	131
5.5	Feature importance for the RF classifier based on the F1-scores for each trust class. The x-axis shows all the PBs, while the y-axis shows the accuracies achieved by each PB as one feature to predict the class of trust.	133
5.6	Comparison of Physiological Behaviours (PBs) Between Competitive and Collaborative Settings	139
5.7	Trust levels across sessions for competitive and collaborative interactions, measured through PBs (EDA, BVP, HR, SKT, BR, BD).	142
5.8	Distrust levels across sessions for competitive and collaborative interactions, measured through PBs (EDA, BVP, HR, SKT, BR, BD).	143

5.9	Learning curve for the Random Forest classifier in the collaborative setting showing training and cross-validation scores as a function of training set size. The minimal gap between training and validation scores demonstrates excellent generalisation capabilities with no overfitting.	144
5.10	Confusion matrix for the Random Forest classifier in the collaborative setting showing the distribution of true positives, false positives, true negatives, and false negatives.	145
5.11	ROC curve for the Random Forest classifier in the collaborative setting showing the trade-off between true positive rate and false positive rate at different classification thresholds. The AUC of 0.87 indicates excellent discriminative ability.	146
5.12	Feature importance for the RF classifier based on SHAP mean value. The x-axis shows all the PBs, while the y-axis shows the SHAP mean value for each PB to predict the class of trust.	148
5.13	Classification Accuracies Using Incremental Transfer Learning . . .	150
5.14	Confusion Matrix for Decision Tree on Target Dataset	151
5.15	ROC Curve for Decision Tree on Target Dataset	152
5.16	Learning Curve for Decision Tree on Target Dataset	152
6.1	Study Procedure	162
6.2	Behaviours Changes Over Time	169
6.3	Top 10 most important behavioural features identified by the Random Forest model	171
6.4	Comparison of training-testing accuracy gaps between initial and regularised models, showing significant reduction in overfitting . . .	172
6.5	Learning curves comparing initial and regularised Random Forest models, showing reduced gap between training and validation scores after regularisation	173
6.6	ROC curves comparing initial and regularised models, showing maintained discriminative ability despite regularisation	174
6.7	Confusion matrices for initial (top row) and regularised (bottom row) models, showing the distribution of predictions across trust and distrust classes	175
7.1	Frozen Lake & Battleship Environment	185

7.2	Learning Curve of Q-learning in Frozen Lake with Robot Advice.	. 189
7.3	Learning Curve of Q-learning in Battleship with Robot Advice.	. . 190
7.4	States Trust Levels for Frozen Lake and Battleship Experiments.	. . 191

List of Tables

2.1	Core Elements of Trust Definitions in HRI	15
3.1	Frequency of factors affecting trust across the three scenarios in study 1(KSA) and study 2(UK).	51
3.2	Mean (M) and standard deviation (SD) of TPS score and relevance score across the three scenarios in Studies 1 and 2. Bold values indicate significantly higher scores in cross-cultural comparisons. Significance levels: $*p \leq 0.05$, $**p \leq 0.03$, $***p \leq 0.001$	52
3.3	Frequency of new factors affecting trust across the three scenarios in study 1 and study 2 identified by the participants.	54
3.4	Frequency of factors affecting human trust in robots across scenarios in Study 1 and Study 2.	55
4.1	Truth table of B_i , C_i and P_i at the i th interaction.	75
4.2	Trust Perception Scale (TPS) 14-item Subscale	80
4.3	Mean (M) and standard deviation (SD) of TMS and TPS scores across the four sessions. The Shapiro-Wilk test confirmed that both TMS ($W = 0.967, p = 0.085$) and TPS ($W = 0.971, p = 0.310$) follow a normal distribution, justifying the use of mean as a central measure.	85
4.4	Truth Table for $ P_i C_i - C_i R_i $	94
4.5	Truth Table for $ K_i - F_i $	94
4.6	Dispositional Trust Questionnaire Items	100
4.7	Means and Standard Deviations (SD) for TPS and TMS across Sessions	102
4.8	Means and Standard Deviations (SD) for TPS and TMS across Sessions	111
5.1	Mean (M) and Standard Deviation (SD) for the physiological features of trust (999 cases) and distrust (480 cases) during all sessions.	125

5.2	Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session.	126
5.3	Classifier Accuracies for physiological Behaviours in Trust Classification.	127
5.4	F1-scores for the five classifiers to predict human's trust and distrust levels. Bold RF is the classifier that achieves the highest accuracy.	128
5.5	Mean (M) and Standard Deviation (SD) for physiological features under trust and distrust conditions for competitive (comp) and collaborative (collab) studies.	140
5.6	Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session in the Competitive setting.	140
5.7	Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session in the Collaborative setting.	141
5.8	Classifier Accuracy's and F1-scores for Physiological Behaviours in Trust Classification in a Collaborative HRI.	147
5.9	Classifier Accuracies and F1-scores for Physiological Behaviours in Trust Classification in a Competitive HRI	147
6.1	List of 51 Facial Blendshape Features	164
6.2	Means and Standard Deviations (SD) for TPS across Sessions	166
6.3	Significant Results from the Mixed-Effects Model Analysis	168
6.4	Mean (M) and Standard Deviation (SD) for the behavioural variables of trust and distrust states during each session.	168
6.5	Accuracy and F1-Scores for regularised Classification Models	170
6.6	Comparison of Initial and Regularised Models: Training-Testing Accuracy Gap and Performance Metrics	172

List of Abbreviations

HRI	H uman R obot I nteraction
HRC	H uman R obot C ollaboration
PB	P hysiological B ehaviour
ML	M achine L earning
RL	R einforcement L earning
MDP	M arkov D ecision P rocess
WoZ	W izard of O z
TPS	T rust P erception S cale
TMS	T rust M odelled S core
EDA	E lectrodermal A ctivity
BVP	B lood V olume P ulse
HR	H ear T R ate
BR	B linking R ate
BD	B linking D uration
RF	R andom F orest
LR	L ogistic R egression
SVM	S upport V ector M achine
DT	D ecision T ree
AB	A da B oost
NN	N eural N etwork
NB	N aive B ayes

Chapter 1

Introduction

The industrial world is experiencing a significant transformation with the emergence of Industry 5.0 [14]. Building upon the automation-driven principles of Industry 4.0, Industry 5.0 strongly emphasises Human-Robot Collaboration (HRC) within flexible and sustainable environments [137]. Rather than perceiving robots only as tools for automation, this new paradigm expects them as collaborative partners that work alongside humans to enhance productivity and produce innovation [14]. In this framework, robots are designed to increase human capabilities, assisting with decision-making, problem-solving, and creative tasks across diverse fields, including manufacturing, healthcare, and personalised services [24, 102, 174, 23]. These robots, ranging from autonomous vacuum cleaners to advanced surgical assistants, not only perform physical tasks but also engage socially with humans, make decisions, and learn from their environments [41]. As robots become increasingly integrated into human environments, the concept of trust plays a crucial role in shaping the dynamics of human-robot interaction (HRI) [92]. In HRI, trust is defined as a multifaceted psychological phenomenon involving beliefs and expectations about a robot's reliability and safety, which develop from experiences and interactions in uncertain and risky situations [1]. Trust is fundamental in both collaborative and competitive settings, shaping the dynamics and outcomes of HRI. In collaborative environments, trust facilitates effective teamwork and coordination between human operators and robotic systems, ensuring smooth task execution and optimising shared goals [194]. Conversely, in competitive contexts, trust dynamics take on a different form, where the robot's truthfulness and adherence to ethical principles play a pivotal role in building and maintaining trust [147]. In both scenarios, trust not only influences human reliance on robotic systems but also determines the overall efficiency and success of the interaction.

1.1 Motivation

The calibration of trust is essential for ensuring successful HRI, as it directly affects the efficiency and effectiveness of interactions between humans and robots. Incorrectly calibrated trust can lead to either over-reliance or under-reliance, potentially resulting in the disuse of advanced robotic systems [157]. To address this, researchers in HRI are exploring the development of an online measurement to sense trust in real-time [92]. The modelling of human trust in robots is integral to the success of HRI as it directly influences the functionality and outcome of interactions between humans and robots [77]. However, it is challenging to measure human trust due to its dynamic nature and the multitude of factors influencing it [91]. Trust evolves dynamically during HRI in various settings through various factors, including reliability, performance, error rates and the context of the interaction [160, 5]. For example, a robot's error rate in a healthcare setting may impact trust differently than in a manufacturing context, emphasizing the need for trust calibration strategies across different environments [18].

Trust in HRI can be understood through its role in both competitive and collaborative settings. In competitive or task-oriented scenarios, truthfulness is critical. It pertains to a robot's ability to provide accurate and transparent information, which is essential for successful interactions where human users rely on the robot's honesty and clarity to make informed decisions [147]. Conversely, in collaborative environments, trustworthiness becomes more significant. This dimension emphasises the robot's dependability and consistent performance over time, fostering a sense of partnership and mutual reliance as humans and robots work together to achieve shared goals [32]. Together, these dimensions of trust—truthfulness in competitive contexts and trustworthiness in collaborative ones—shape how humans develop and adjust their trust in robots across various HRI scenarios.

In HRI, researchers evaluate human trust using two primary methods: subjective and objective trust measurements [91]. Subjective measurements, the more commonly used approach, rely on participants' responses to questionnaires to assess their perceptions and feelings of trust towards robots [53]. Objective methods, in contrast, focus on analysing participant behaviour during interactions, providing data-driven insights into trust that are not dependent

on self-reported perceptions [106]. While subjective methods are effective for capturing users' opinions and emotional states, they often fail to reflect the dynamic, real-time nature of trust during interactions [22]. Objective methods offer critical behavioural insights, enabling robotic systems to adapt communication strategies and optimise decision-making in real-time [132]. However, objective methods face notable challenges, including difficulties in standardising measures across diverse robotic systems, limitations in addressing the multidimensional nature of trust, and a lack of scalability for practical, real-world applications [108]. Combining subjective and objective methods provides a more comprehensive and reliable framework for measuring trust dynamics. Subjective methods capture users' perceptions, while objective methods provide quantifiable data that bridges the gap between perceived trust and actual behaviour. This dual approach enhances the robustness and validity of findings by identifying discrepancies between self-reported trust and observed behaviour, offering deeper insights into trust dynamics and enabling a more comprehensive estimation of trust that accounts for factors such as interaction duration and the robot's overall performance, as highlighted by Law and Scheutz [106].

Building upon the strengths of subjective and objective methods, the development of mathematical models of trust in HRI has emerged as an important avenue of research [64]. These models aim to integrate insights from both approaches to provide structured frameworks for quantifying and predicting trust. By leveraging data from past interactions, mathematical models can track trust transfer across tasks and adjust trust dynamically during repeated and long-term HRIs [3, 195, 65]. However, capturing the essence of trust in HRI remains a challenging task [195]. Several studies have attempted to create real-time mathematical models of trust [55, 87, 80]. However, these models are not without limitations. One primary issue is that the validation of these models has occurred in simulated environments, which raises questions about their practicality in real-world HRI scenarios [92]. Moreover, these models lack generalisation across different settings and fail to account for the variations in factors influencing trust in different environments.

Existing approaches, such as mathematical models, provide a structured framework for estimating and calibrating trust. However, there is a growing

need to complement these models with methods that can account for the dynamic, real-time, and often subconscious nature of trust-related emotional and cognitive responses during interactions. One promising approach is the use of physiological behaviours (PBs), which builds upon the concept of objective trust measurement [144]. PBs, such as electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), blinking rate (BR) and blinking duration (BD), offer a promising avenue for understanding and assessing trust in real-time [7]. These behaviours indicate an individual's emotional and cognitive states during interactions with robots. By monitoring these responses, we gain valuable insights into the dynamic nature of trust and its influence on HRI [6]. Variations in physiological signals are observed across both trust and distrust states, with these states assessed by correlating the physiological data with participants' behavioural decisions during tasks specifically designed to elicit varying levels of trust. For instance, in HRI scenarios, distrust may be reflected in heightened HR and EDA due to increased cognitive effort and emotional arousal. Researchers have conducted several studies on assessing human trust in robots by using physiological indicators. However, these studies have limitations. Most notably, they used a limited number of physiological indicators in combination. Additionally, most of these studies were conducted in simulated environments rather than in real HRI settings. Moreover, existing research has mainly focused on collaborative settings.

The use of facial emotions and voice features to predict human trust in robots is another innovative and promising approach [91]. Facial expressions and vocal tones provide real-time data that reflect a person's emotional and cognitive states during interactions with robots [49, 90]. However, there has been limited research into detecting trust from user speech and facial expressions [89, 60]. They demonstrate the potential of using facial expressions and speech features to estimate trust, though challenges like dataset bias and modest accuracy highlight the need for further refinement. Additionally, there is a lack of comprehensive studies that integrate these multi-modal cues to predict human trust in robots in real-time. Most existing research tends to focus on either facial emotion and voice features in isolation, rather than combining them with facial remarks to provide a complete understanding of how these cues together influence trust. In this work, our aim is to integrate vocal features, facial emotion, and facial

markers to predict human trust in robots.

The importance of measuring trust during long-term or repeated HRI has been recognised, but there is limited work in this area [125]. Trust calibration is essential in long-term and repeated HRI, as it involves adjusting trust levels based on the robot's performance, human experience, and evolving expectations [42]. In this context, research in HRI indicates that trust levels change according to experiences gained over time [78, 39]. To the best of our knowledge, there has been no prior examination of modelling human trust in robots and analysing physiological behaviours to assess and monitor trust in repeated interactions. This thesis aims to address this gap by mathematically modelling human trust in robots and utilising physiological behaviour, facial expressions, and vocal cues in repeated interactions.

Another important aspect is how reinforcement learning (RL) can be utilised to optimise human interaction with robots. RL has the potential to dynamically adapt robot behaviours based on continuous feedback from human users, thereby enhancing the quality and efficiency of HRI [197]. Despite its potential, its application specifically for optimising trust in HRI, particularly using validated metrics and in the context of repeated interactions, remains limited, leaving a critical gap in the development of adaptive and trustworthy robotic systems [2].

To address these challenges, this thesis focuses on developing an integrated framework that combines mathematical modelling, physiological behaviour analysis, and reinforcement learning to advance real-time trust measurement and optimisation in HRI. By enhancing the accuracy and applicability of trust models, this research aims to support the development of adaptive and reliable robotic systems capable of repeated and long-term interactions. Rigorous validation in real-world settings ensures the creation of efficient, trustworthy, and human-centric robotic systems.

1.2 Research Objective

In this thesis, we aim to develop and evaluate a model and measure of human trust in robots during repeated interactions. Repeated interaction is a central

focus of this research as it allows us to observe how trust evolves over time and across multiple encounters with robotic systems. The core objectives of this research are as follows:

- Develop and validate a computational model to capture the dynamic evolution of human trust in real-time during repeated HRI, examining how trust changes across multiple interaction sessions in different settings.
- Explore the use of physiological indicators such as heart rate and electrodermal activity to measure and track changes in human trust in robots throughout repeated interaction sessions, enabling real-time trust assessment across different contexts.
- Investigate how vocal and non-vocal cues, including facial expressions, facial movements, and voice features, evolve during repeated interactions with robots, and utilise these patterns to assess and predict human trust development over time.
- Integrate reinforcement learning techniques with trust models to optimise trust measurement and improve robotic system performance in simulated environments.

To achieve these research objectives, we expect to find answers to the following research questions (RQs), each directly aligned with our research objectives:

- **RQ1:** What are the most important factors affecting user trust in robots across various HRI settings? [Chapter 3]
This question supports our first objective by identifying the key factors that must be incorporated into our computational model of trust during repeated interactions.
- **RQ2:** How can we model and validate human trust in robots during repeated HRI in different settings? [Chapter 4]
This question directly addresses our first objective of developing a computational model for trust in repeated interactions.
- **RQ3:** How can human physiological behaviours be used to predict trust levels in robotic agents throughout different repeated interactions? [Chapter 5]

This question aligns with our second objective of using physiological indicators to track trust changes during repeated HRI.

- **RQ4:** How can human vocal and non-vocal cues be used to predict human trust in robots during repeated interactions? [Chapter 6]

This question corresponds to our third objective of investigating vocal and non-vocal cues as they evolve during repeated HRI.

- **RQ5:** How can reinforcement learning be utilised to optimise human trust in robots during HRI? [Chapter 7]

This question addresses our fourth objective of integrating reinforcement learning with trust models in simulated environments.

To address these research questions, we started with an initial study designed to explore the factors that influence human trust in robots across various settings and cultures. This study provided valuable insights into the most important factors affecting trust. Based on these findings, we developed our initial mathematical model, focusing on experience-based performance and user controls as key elements. We tested this model in a repeated competitive HRI scenario, and the results generally confirmed its validity. However, our analysis revealed that the Trust Prediction Score (TPS) alone was not a reliable predictor of trust. Additionally, we found that the level of risk varied across different sessions. Building on these insights, we refined the model by incorporating the critical factors identified and tested in a different context, as suggested by the results in Chapter 3. Therefore, we tested this improved model in two collaborative HRI settings, which again validated its accuracy and showed it to be superior to the initial version. Additionally, we explored an alternative approach to measuring human trust in robots by predicting trust levels based on physiological behaviour. For this purpose, we carried out two separate user studies in different settings: one focused on truthfulness (competitive) and the other on trustworthiness (collaborative). These studies aimed to evaluate the effectiveness of physiological indicators such as electrodermal activity, heart rate, blood volume pulse, and other relevant metrics in assessing trust in real-time interactions. Moreover, we investigated how human vocal and non-vocal, including facial emotions, facial movements, and voice patterns, can be used to predict human trust in robots in real-time during repeated HRI. The study showed that these signals effectively predict trust levels with an accuracy rate

of 77%. Finally, we utilised our validated model to create a proof of concept for a robotic system that can adapt to different situations. We developed a reinforcement learning application in a simulated environment that leverages the mathematical model to build trust between humans and robots. The results demonstrated that the RL model effectively balanced trust by dynamically adjusting it according to task outcomes, thereby improving task performance and minimising the risks associated with under and over-trust.

By systematically addressing these objectives, this research has significantly advanced the understanding and optimisation of human-robot trust dynamics, contributing to the development of more reliable and adaptive robotic systems.

1.3 Contributions

The following are the main contributions to the field of HRI from the work presented in this thesis:

Contribution 1: Our first contribution, addressing RQ1, lies in providing a comprehensive analysis of the most important factors affecting user trust in robots in diverse HRI contexts. We identified both common and new factors, such as controllability, familiarity, and risk, which had not been fully explored in previous literature. These findings are essential for guiding the development of our computational trust model, enabling us to incorporate context-specific factors that influence trust. By understanding the contextual differences that impact trust in robots, we ensured that our model accurately captures the dynamic nature of trust across various environments (Chapter 3).

Contribution 2: Our second contribution, addressing RQ2, lies in the development of a computational model designed to calibrate trust dynamically during HRI. This model integrates critical factors such as experience-based performance, user controls, risk and ambiguity aversion, allowing the robot to adjust its behaviour based on real-time feedback. We evaluated the model through experimental tasks in both truthfulness (competitive) and trustworthiness (collaborative) settings, demonstrating its effectiveness and robustness in predicting trust levels and adapting to different HRI contexts.

Refining the model through empirical data ensured that it provides a reliable and robust tool for enhancing trust in repeated HRI (Chapter 4).

Contribution 3: Our third contribution, addressing RQ3, lies in exploring the use of multiple physiological indicators, including EDA, BVP, HR, SKT, BR, and BD, to predict trust levels during repeated HRI. Through extensive user studies conducted in both competitive and collaborative settings, we demonstrated the significant role of PBs such as HR and SKT in distinguishing between trust and distrust states. Furthermore, we were the first to employ incremental transfer learning, which allowed for the adaptation of trust models across different HRI contexts, enhancing the model's capability to generalise trust predictions from one setting to another, achieving an accuracy of 89%. This contribution offers a novel approach to measuring trust by leveraging PBs and providing deeper insights into the physiological basis of trust during HRI (Chapter 5).

Contribution 4: Our fourth contribution, addressing RQ4, lies in investigating how human facial expressions and voice features can be used to predict trust levels in robots during repeated interactions. Through experimental studies involving a collaborative game-based HRI, we analysed the predictive power of vocal cues (such as pitch and harmonicity) and non-vocal cues (such as facial emotions and movements) to classify trust and distrust states. We demonstrated that emotions like fear and anger, facial blendshapes like cheek squint and jaw movement, and vocal characteristics like duration and harmonicity can significantly predict trust levels. The Random Forest classifier achieved a classification accuracy of 77%, which refers to the proportion of correct predictions out of all predictions made by the model. This metric is complemented by other performance indicators such as F1-score and confusion matrices to account for class imbalance and ensure fair evaluation. The accuracy score here reflects the model's ability to generalise and reliably distinguish between trust and distrust decisions across participants. This contribution introduces a new dimension to trust prediction in HRI by utilising multi-modal data, combining vocal and non-vocal cues to enhance real-time trust assessment in robotic systems (Chapter 6).

Contribution 5: Our fifth contribution, addressing RQ4, lies in integrating our validated computational trust model with reinforcement learning techniques to dynamically adjust robot behaviours and optimise human-robot trust. We

conducted experiments in simulated environments (Frozen Lake and Battleship), demonstrating how the RL application balanced trust by dynamically calibrating it based on task outcomes. The results showed that the RL model improved both task performance and trust alignment, reducing the risks of insufficient or extreme trust. This contribution presents a mechanism for real-time trust optimisation in HRI, highlighting the potential for RL to enhance human-robot collaboration by offering practical solutions for trust calibration in real-world scenarios (Chapter 7).

Contribution 6: This thesis involves sharing comprehensive study materials and diverse datasets with the academic community to support ongoing and future research in HRI. The datasets include detailed annotations of physiological behaviours collected from experiments conducted in both competitive and collaborative HRI settings. Additionally, another dataset includes vocal and non-vocal cues, including facial expressions, facial movements, and voice patterns, providing a multi-modal perspective on trust measurement. We also offer experimental materials, including tasks and machine-learning scripts. The datasets and materials can be accessed [here](#).

1.4 Thesis Outline

The remaining chapters of this thesis are structured as follows: In Chapter 2, we review the relevant literature on trust in HRI, including trust measurement techniques and factors influencing trust. Chapter 3 presents the findings from our initial study on the key factors affecting trust in various HRI settings. Chapter 4 discusses the development and validation of the computational trust model. In Chapter 5, we explore the use of physiological indicators in predicting trust levels, while Chapter 6 examines the role of vocal and non-vocal cues in trust prediction. Chapter 7 details the integration of reinforcement learning techniques for optimising human-robot trust. Finally, Chapter 8 summarises the contributions of this thesis and outlines future research directions.

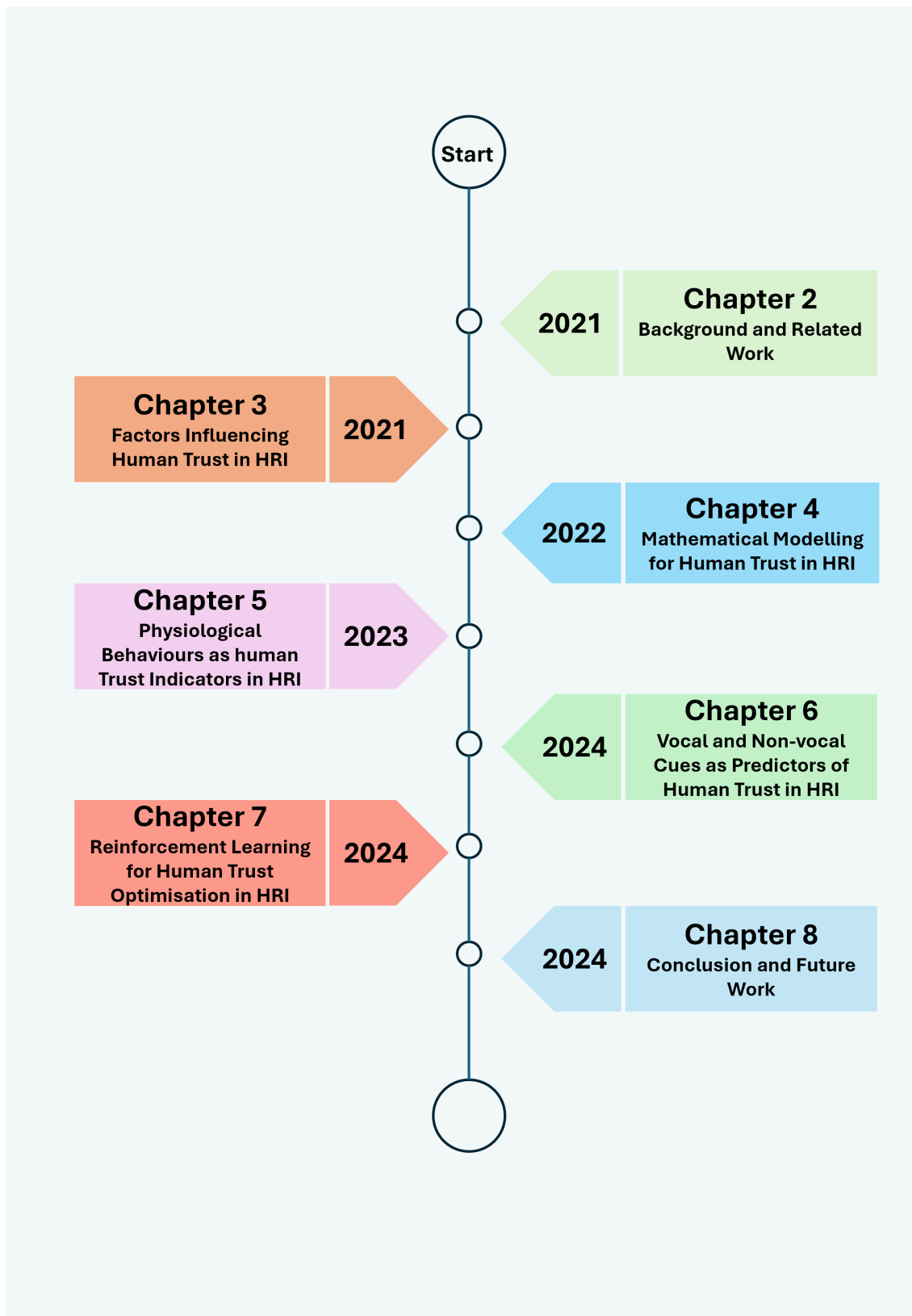


FIGURE 1.1: Thesis Road-map

Chapter 2

Review of Related Work

This chapter provides a review of the related work connected with the research presented in this thesis, which aims to investigate how to model and measure human trust in robots in real-time during human-robot interaction (HRI). The research examines various methods for modelling and measuring human trust in robots, as well as strategies for optimising human trust in robots. Section 2.1 offers an overview of the theoretical understanding of trust in general and in the field of HRI. Section 2.2 provides an overview of the factors influencing human trust in robots, including trust transfer across settings and cultures. Section 2.3 includes a comprehensive review of studies that measure human trust in robots using subjective, objective, physiological behaviour, facial expression, and voice features. Section 2.4 provides an overview of studies focused on long-term and repeated interactions, examining how trust dynamics evolve over time during HRI.. Section 2.5 covers studies that have attempted to use reinforcement learning to optimise human trust in robots.

2.1 Theoretical Understanding of Trust

Trust is a multidimensional concept that has gained significant attention in research across various fields for decades, including organisational behaviour, psychology, cognitive science, human-computer interaction, and HRI [70, 6, 93]. From a physiological perspective, trust is a human's mental state. It extends beyond interpersonal interactions and influences various forms of interactions, including those between humans and organisations, as well as between humans and technologies. While numerous studies have explored trust, its definition varies across fields due to the complexity and context-specific nature of trust

[91]. This variation reflects the multifaceted nature of trust and underscores the importance of studying trust in specific domains like HRI, where unique factors such as robot autonomy, uncertainty, and vulnerability come into play.

There has been considerable effort in the domain of HRI to conceptualise trust. The definition of trust may vary based on the robot's application and domain [92]. To date, more than 34 different definitions of trust have been developed. While it is challenging to evaluate trust between humans and robots based only on this diverse array of definitions, there are several common characteristics shared among these definitions that are essential for understanding trust across various domains [158, 71].

To establish a clear understanding of trust in HRI, we identify and explore three essential dimensions that emerge from the literature: *Cognitive Expectations*, *Physiological Manifestations*, and *Contextual Dynamics*. Each of these dimensions plays a critical role in shaping how trust is formed, maintained, and measured in HRI.

Cognitive Expectations: Trust as a Mental Construct

Trust often begins as a mental state where the trustor (human) forms expectations about the trustee (robot) based on prior experience or perceived reliability. Cognitive expectations are foundational in determining whether an individual will engage with a robot and to what extent they are willing to rely on the robot's actions.

Mayer, Davis, and Schoorman [120] define trust as the “willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor.”

This definition places *expectations* at the core of trust, especially in situations where the trustor has little control over the trustee's actions.

Rotter [150] highlights a similar cognitive aspect, describing trust as a “generalised expectancy that the word or promise of another can be relied on.”

In HRI, these *cognitive expectations* are critical in shaping how humans perceive robot reliability, competence, and autonomy. Whether in collaborative tasks or autonomous operations, trust is fundamentally linked to how humans cognitively evaluate the robot's ability to meet their goals.

Physiological Manifestations: Trust as a Subconscious Response

While trust begins as a cognitive process, it is often expressed and experienced subconsciously through physiological signals. Trust, particularly in uncertain or high-risk situations, can induce emotional and physiological responses such as changes in heart rate, skin conductance, and even facial expressions.

Ajenaghughrure, Sousa, and Lamas [7] define trust as a “subconscious cognitive process that involves mental reasoning, memory, and learning”, noting that physiological indicators offer a window into real-time trust dynamics.

Contextual Dynamics: Trust as Situational and Adaptive

Trust in HRI is highly context-dependent, and the nature of the task or interaction significantly influences how trust is established, maintained, or lost. Trust may vary depending on whether the interaction is collaborative, where the human and robot are working together toward a common goal, or competitive, where the robot's ability to provide accurate, truthful information is crucial.

Lee and See [107] emphasise the importance of context in their definition of trust as an “attitude that an agent will help achieve an individual's goals in a situation characterised by uncertainty and vulnerability.”

Hancock et al. [70] similarly propose that trust in robots is the “reliance on a robot to perform a task or role, especially in environments where human control is limited or absent.”

While trust is a multidimensional construct, certain elements consistently appear across the aforementioned definitions. These elements are critical for understanding, modelling and measuring human trust in real-time during HRI.

Core Element	Explanation
Expectations	Trust involves the belief that the robot will perform actions aligned with the trustor's goals and expectations, particularly in uncertain situations.
Vulnerability	Trust requires the willingness to be vulnerable, where the trustor depends on the robot's actions despite potential risks or uncertainty.
Risk and Uncertainty	Trust is crucial in environments characterised by risk and uncertainty, where the outcome depends on the robot's performance and decisions.
Reliance	Trust reflects the reliance on a robot to perform tasks autonomously or to provide truthful and accurate information in critical situations.
Physiological Response	Trust manifests subconsciously through physiological indicators like heart rate, EDA, and facial expressions, offering real-time insights into trust.
Context Dependence	Trust varies across different interaction contexts (e.g., collaborative vs. competitive), shaping the dynamics of HRI.

TABLE 2.1: Core Elements of Trust Definitions in HRI

Building on these definitions and core elements of trust in table 2.1, this thesis develops a comprehensive framework for real-time trust measurement in HRI. By integrating cognitive expectations, physiological responses, and contextual dynamics, the framework aims to dynamically assess and calibrate trust during HRI.

2.2 Factors Affecting Human Trust in Robots

While Hancock et al. [71] identified a broad range of factors influencing trust in HRI, their model does not sufficiently address the critical roles of risk and vulnerability in shaping human trust during interactions with robots. Research has shown that trust dynamics can vary significantly depending on the level of risk associated with the task or environment [145]. In high-risk contexts, such as emergency response scenarios, the potential for negative outcomes boosts the importance of proper trust calibration. Additionally, Abbass, Scholz, and Reid [1] have emphasised the need for the HRI community to explore context-specific characteristics, including uncertainty and vulnerability, with a particular focus

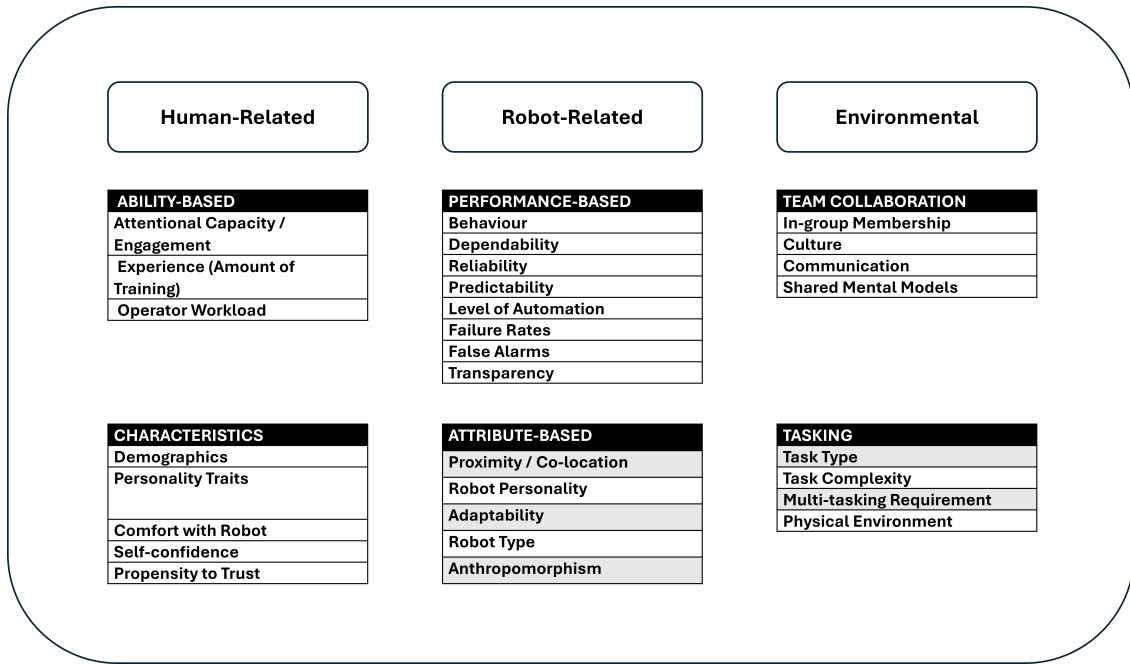


FIGURE 2.1: Hancock's three factors of trust model

on the potential gains or losses for the trustor. This aspect becomes crucial in high-stakes environments.

Although Hancock's framework provides a valuable starting point, it overlooks the significant impact of vulnerability, a concept central to many trust definitions in HRI. As highlighted in a recent meta-analysis by McKenna et al. [123], vulnerability arises when users depend on robots in uncertain situations, particularly in tasks where the robot's performance or reliability is questionable. The study suggests that in such contexts, trust-building is heavily dependent on enhancing the robot's task performance. This gap in Hancock's model emphasises the need for further exploration of trust formation in various situations involving different perceived vulnerabilities and risks.

Moreover, Hancock's framework does not account for cultural differences, which play a critical role in shaping trust in HRI. Cross-cultural variations in expectations of robot behaviour, uncertainty, and perceptions of risk can significantly influence how trust is developed, as demonstrated by recent studies [156, 50]. Therefore, in this thesis, we extend Hancock's model by investigating more factors influencing trust, such as risk and vulnerability and how trust varies across settings and cultures.

2.2.1 Trust transfer across settings

Various factors influence humans' trust in robots, and these factors can vary across different environments [99]. The intensity of each factor may be more evident in certain contexts depending on the nature of the task, such as the rationality and reversibility of the task [145], risk to human safety [153], or the workload [43]. For example, trust may be impacted more in situations where a robot operates in close proximity to a human, as the perceived risk can be more significant in such cases [110]. Despite the significance of these factors, the phenomenon of trust transfer from one context to another still needs to be explored in HRI research. A few studies have investigated this directly, such as [145, 167, 184, 123].

Robinette et al. [145] suggested that trust factors may vary across different domains and environments due to the level of risk. They conducted an experiment in HRI in a non-emergency task to observe human behaviour and then decide whether or not to follow the robot's instructions in an emergency evacuation scenario. Their results showed that the trust varied between emergency and non-emergency scenarios [145]. However, this work focused only on the risk level and did not compare tasks with the same level of risk. Soh et al. [167] investigated trust variations in robots across household and driving tasks, noting that trust changed depending on the specific nature of the task. Similarly, Xie et al. [184] explored the role of robot capabilities and intentions in shaping trust across tasks like searching, mapping, and firefighting. They found that trust shifted based on the robot's abilities and performance in different tasks. While these studies demonstrated that task-specific factors influence trust transfer, they largely focused on robot performance without considering other dimensions, such as user vulnerability or environmental constraints. Wong, Xu, and Dudek [182] explored trust in shared control tasks using drones. Their findings demonstrated that trust was influenced by both task complexity and user vulnerability, reinforcing the idea that trust dynamics change across different contexts depending on the amount of control users feel they have and the risks they perceive. Sebo, Krishnamurthi, and Scassellati [164] found that when robots displayed vulnerable behaviours, such as expressing uncertainty or hesitancy, human teammates tended to trust them more. This suggests that a robot's perceived vulnerability, along with the inherent risks of the task,

plays a crucial role in trust transfer across different settings. For example, in industrial environments where robots are seen as highly capable but potentially intimidating, trust may not transfer as easily to lower-risk settings like domestic environments, where robots are perceived as less physically imposing and more dependent on human input. More recently, McKenna et al. [123] conducted a meta-analysis on trust in HRI, identifying a range of factors that influence trust across various settings. They categorised vulnerability into four key dimensions: task-related vulnerability, privacy concerns, financial risk, and physical risk, each of which impacted trust differently depending on the situation. For instance, studies involving high-risk settings where users relied on robots to perform safety-critical tasks (task-related vulnerability) showed that trust was highly sensitive to robot performance [123]. This suggests that trust transfer across different settings may depend not only on the robot's capabilities but also on the perceived risk and specific vulnerabilities associated with the task.

This research expands on existing studies by examining a wider range of factors that shape trust across various settings. This broader perspective on trust transfer sets the stage for an exploration of how cultural differences further influence trust dynamics in HRI.

Trust in Human-robot Competition

In exploring trust transfer across different settings, it becomes clear that competitive interactions introduce unique challenges compared to traditional collaborative contexts in HRI [193]. Most existing research on trust in HRI has been conducted in collaborative settings, where humans and robots share common goals and success metrics. These studies highlight how trust depends on the robot's reliability, safety, and alignment with human objectives [71]. However, in competitive interactions, where humans and robots directly compete against each other for rewards or task outcomes, the dynamics of trust transfer are altered significantly [191].

In competitive settings, trust is influenced by additional factors, such as the robot's truthfulness and perceived competence relative to the human. For instance, Sebo, Krishnamurthi, and Scassellati [164] demonstrated that a robot's honesty in fulfilling promises during competitive tasks impacted human trust levels, especially when the robot broke a promise in a competitive scenario. This

highlights how truthfulness, a factor less prominent in collaboration, becomes critical in competition, where distrust can arise quickly if the robot appears deceptive or untrustworthy.

Furthermore, emotional responses play a significant role in competitive HRI. Kirtay et al. [95] found that trust in competitive settings is closely tied to emotional responses like frustration or satisfaction. Positive emotions, such as contentment when the robot performs well, tend to enhance trust, whereas negative emotions, like frustration when the robot fails or competes too aggressively, can diminish it. This suggests that competitive settings evoke different emotional and cognitive processes, making trust transfer more challenging from collaborative to competitive settings.

To bridge this gap, the current thesis examines trust dynamics in both collaborative and competitive settings, investigating how factors like robot truthfulness, emotional impact, and task-specific performance influence trust transfer. By expanding the focus beyond traditional collaborative contexts, this work provides insights into the unique demands of competitive interactions in HRI, offering a more comprehensive view of trust dynamics across diverse settings.

2.2.2 Culture and trust

A growing body of research has explored the influence of cultural background on trust in human-robot interaction (HRI) (e.g., [140, 178, 131, 112]). These studies consistently demonstrate that individuals from different cultures exhibit varying levels of trust towards robots. However, most cross-cultural HRI research has focused on comparisons between Western and East Asian cultures, with only limited attention given to other regions, including the Arab world [71]. The Arab region, encompassing 22 countries and home to over 620 million people across Asia and Africa [165], presents a significant yet underexplored area in the study of HRI. Given the cultural and contextual diversity of the Arab world, it is essential to include this region more comprehensively in future HRI research.

For example, Rau, Li, and Li [140] examined how cultural background, robot appearance, and task type affected trust, likeability, and engagement in participants from China, Korea, and Germany. They found significant

cultural differences in trust and engagement, indicating that trust in robots is shaped not only by the robot's appearance and task but also by the user's cultural background. Similarly, Nomura, Syrdal, and Dautenhahn [131] used the *Frankenstein Syndrome Questionnaire* [173] to assess social acceptance of humanoid robots in Japan and the UK, finding notable differences in the acceptance of robots between these two cultures. However, this study was limited to humanoid robots, leaving open questions about how trust transfers across different robot types in varied cultural contexts.

In a more culturally specific study, Andrist et al. [20] compared perceptions of robot credibility between Arabic-speaking participants in Lebanon and English-speaking participants in the United States during a collaborative task. The results showed that participants from Lebanon were more critical of robots' perceived credibility compared to their American counterparts. This suggests that cultural factors, including language, may influence how robots are perceived and trusted. More recently, Lim, Rooksby, and Cross [112] reviewed global cultural influences on HRI, examining factors such as knowledge, expectations, perceptions, and behaviours toward robots. Their findings reinforced the idea that cultural context significantly affects how robots are trusted and accepted across different societies.

In summary, previous research highlights the role of culture in shaping trust in HRI, but there remain several gaps in the literature. Most notably, there is a lack of comprehensive studies investigating trust dynamics in diverse cultural settings, particularly in regions like the Arab world. Furthermore, much of the existing research that considers Arab culture focuses on language as a primary cultural factor, overlooking other important aspects of culture, such as social norms, traditions, and values. This thesis examines how trust in robots varies across cultures, focusing on the dynamics in various HRI settings and with different types of robots.

Section's 2.2 Gap:

Existing research has explored various factors influencing human trust in robots, but it remains limited in addressing how trust dynamics shift across different settings, particularly concerning risk and vulnerability. Additionally, there is still much to understand about how cultural differences affect trust in robots and the specific factors that influence trust within each culture. This thesis builds on prior work by examining how these elements impact trust across diverse environments and cultural contexts, providing a more comprehensive understanding of trust in HRI.

2.3 Measuring Human Trust in Robots

Researchers have developed various methods to measure human trust in robots, each with its own strengths and limitations. A commonly employed approach involves the use of questionnaires, where participants are asked to evaluate their level of trust in a robot based on a range of factors, such as the robot's perceived usefulness, reliability, accuracy, and ease of use [21]. This subjective assessment of trust can provide valuable insights into the user's overall perception of the robot's trustworthiness. They made an effort to create more widely applicable measures [84, 189, 159, 119], and some of these are beginning to be more frequently used in HRI trust studies.

In addition, researchers have also recognised the importance of going beyond self-reported measures. This may include observing the user's interactions with the robot, such as their willingness to follow the robot's recommendations or to allow the robot to perform tasks autonomously [91, 106]. As part of this objective measurement of trust during HRI, efforts have been made to mathematically model humans' trust in robots during HRI [55, 79, 88, 186, 152, 103, 69] in collaborative contexts.

2.3.1 Mathematically modelling

Freedy et al. [55] proposed a Collaborative Performance Model for human-robot collaboration that emphasised the importance of trust in improving team performance in mixed-initiative environments. The model categorised trust into three levels: under-trust, proper trust, and over-trust, based on

operator interventions and robot competency. Using the MITPAS simulation environment, the study tested the impact of robot autonomy on operator trust and performance. Participants interacted with unmanned ground vehicles (UGVs) in a military simulation and were responsible for managing the UGV's targeting and firing tasks at varying competency levels. Results indicated that operators compensated for low robot competency by increasing manual control, and trust levels fluctuated based on the perceived reliability of the robots. The findings underscored the importance of appropriate trust calibration for effective human-robot collaboration. However, the study has limitations, including a small sample size and the use of a simulated environment.

Hoogendoorn et al. [79] developed and validated models to capture biased human trust in multi-agent systems. The study focused on how positive, neutral, and negative biases affect human trust evolution over time when humans rely on external advice from agents in decision-making tasks. The authors extended traditional trust models by incorporating cognitive biases and tested six variations of biased trust models, including both linear and logistic forms: linear model with biased experience (LiE), linear model with biased experience influenced by current trust (LiET), linear model with bias based only on trust (LiT), logistic model with biased experience (LoE), logistic model with biased experience influenced by current trust (LoET), logistic model with bias based only on trust (LoT). In the experimental setup, 16 participants completed classification tasks based on simulated UAV video footage, where they had to trust advice from other participants or software agents. The study showed that incorporating biases into the trust models significantly improved prediction accuracy compared to unbiased models. Specifically, the LiE and LoET models outperformed others, indicating that biases play a crucial role in shaping human trust in autonomous systems. However, the study was limited by a small sample size and relatively simple simulated experimental tasks, suggesting the need for more diverse and complex scenarios to further validate the findings.

Kaniasarasu et al. [88] investigated the alignment of user trust with robot performance in real-time during HRI. The study focused on ensuring that users' trust in a robot's abilities matched its actual reliability, thus preventing under-trust or over-trust. The researchers introduced semantic and non-semantic feedback in the form of a robot confidence indicator displayed on a user interface

to help participants adjust their trust during a slalom course task with varying robot reliability. Trust was measured every 20 seconds, with participants allowed to switch between fully autonomous and assisted modes. The study found that providing feedback, either semantic or non-semantic, significantly reduced trust mismatch compared to no feedback. Without feedback, participants tended to over-trust the robot, but with feedback, they better aligned their trust with the robot's performance. Trust alignment scores showed that feedback helped participants recognise when the robot was unreliable, preventing inappropriate trust shifts. In summary, trust mismatch was based on the sum of control taken during an interaction.

Xu and Dudek [186] presented a model called OPTIMo (Online Probabilistic Trust Inference Model) to quantify the degree of trust that a human supervisor has in an autonomous robot over time. The objective was to create a real-time, personalised trust model that adapts based on observed interaction experiences, specifically in asymmetric human-robot collaboration settings where the human occasionally intervenes in the robot's tasks. The method involved constructing a Dynamic Bayesian Network to infer human trust based on robot performance, human interventions, and additional factors. The testing was carried out using data from an observational study involving 21 roboticists who interacted with an autonomous boundary-tracking aerial robot in a simulated environment. The results showed that OPTIMo could accurately predict trust levels in real-time, outperforming several existing models in terms of prediction accuracy and responsiveness. However, limitations of the study included the reliance on a specific group of users (roboticists), a focus on a single task (boundary tracking), and the absence of testing across more diverse and generalised HRI scenarios.

Saeidi and Wang [152] utilised a trust and self-confidence-based autonomy allocation model aimed at reducing human cognitive workload and improving overall system performance in human-robot collaboration systems. The model is grounded in the idea that the difference between a human's trust in the robot and their self-confidence in manually controlling the robot determines how autonomy is allocated between the human and the robot. The method incorporates real-time measures of human and robot performance to develop an objective, quantitative model (TSC) that suggests optimal autonomy allocation. The study involved simulations to evaluate the model's impact on long-term

robot performance and human workload. A case study showed that the TSC-based allocation system successfully captured human behaviour in autonomy allocation and provided a mechanism to gradually correct suboptimal human autonomy decisions through a Nonlinear Model Predictive Control (NMPC) algorithm. The results indicated that the model improved robot performance and balanced human utilisation. However, the study was limited to simulated environments, and further real-world validation in more diverse tasks would be needed to confirm its broader applicability.

Kumar and Dubey [103] developed a cognitive trust model for evaluating robot performance in task-oriented scenarios by focusing on two key factors: capability, which includes the robot's capacity and expertise, and intention, which comprises the robot's desire and commitment to perform tasks. The model was tested using a customised robotic arm (VPL-RAT Robotic Arm Manipulator) that performed path-planning tasks with three different algorithms: an Artificial Neural Network (ANN), a reinforcement-based ANN, and the Situation-Operator Model (SOM). The trust model showed that robot trust increased over five cycles of learning, with the SOM algorithm outperforming the others. However, the study is limited to simulation and would require validation in real-world applications and more complex task environments.

Hu et al. [81] introduced a quantitative model to capture human trust dynamics in human-machine interactions, focusing on how factors like past experience, cumulative trust, and expectation bias influence trust. The study also explored the effects of demographic factors such as national culture and gender on trust behaviour. Two experiments were conducted in a simulated autonomous driving task where participants had to trust or distrust an obstacle detection sensor's reports. The experiments introduced varying reliability, with some trials featuring system misses or false alarms, to elicit dynamic changes in trust. Results showed that the proposed model accurately captured human trust dynamics, achieving over 92% fit accuracy. U.S. participants were found to trust the system less than Indian participants and were more sensitive to system misses, while male participants were more affected by misses compared to female participants. The model highlighted the importance of expectation bias and cumulative trust in shaping trust over time. However, the study's reliance on a computer-based simulation limits its validity.

Hale, Setter, and Fregene [69] developed a trust model that adapts human users' privacy levels in HRI based on the robot's cooperation over time, utilising differential privacy mechanisms. The model incorporated variables such as trust propensity, trust decay, and robot cooperation, mapping these to dynamic privacy controls. The study used simulations to demonstrate how robots cooperating with human users over time could increase trust, leading to reduced privacy noise. Conversely, when a robot stopped contributing to a decrease in cost, trust decayed, increasing the noise in the data shared with the robot. While the model effectively captured trust dynamics and privacy relationships, the study was limited to simulations, lacking real-world validation, and did not consider more complex human behaviour or environmental factors that could arise in practical HRI contexts.

Section's 2.3.1 Gap:

In summary, current mathematical models of trust in HRI have largely focused on aligning human trust with robot performance and autonomy, often overlooking critical factors such as risk and uncertainty. These models have primarily been developed and tested in simulated environments, with an emphasis on collaborative contexts. While models such as those proposed by Freedy et al. [55] and Hoogendoorn et al. [79] offer valuable insights into trust calibration, their validation across different settings or through physical robot interactions remains limited. This thesis addresses these gaps by proposing and testing the model that integrates crucial factors influencing trust dynamics, including risk and uncertainty. In addition, these models are validated through physical interactions across three different settings, offering a more comprehensive understanding of trust in human-robot interactions.

2.3.2 Physiological Behaviour Measurement

While the mathematical models discussed earlier, provide a structured and reliable framework for estimating and calibrating trust. However, to further enhance their validity and applicability, there is a growing need to complement these models with methods that can capture the dynamic, real-time, and often subconscious nature of trust-related emotional and cognitive responses during interactions. Therefore, the researcher proposed physiological behaviours (PBs) as alternative methods for assessing trust in HRI [29]. The physiological

behaviour of the human body is a complex interplay of various biological systems, including the nervous and endocrine systems, along with their associated processes [127]. These systems ultimately shape and influence human behaviour, cognition, and emotional experiences [44, 96]. To monitor and track these responses, the human body has four distinct organs that constantly record changes in physiological sensors. The brain, the most complex organ in the body, measures neurological activity through an electroencephalogram (EEG). The heart, a powerful muscle that pumps blood throughout the body, measures HR or BVP. The skin, the body's largest organ, measures EDA or SKT. Finally, the eyes, the body's primary sensory organs, measure BR or BD.[4]. 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 [94].

Empirical Evaluations and Physiological Behaviours

PBs have been studied across various disciplines, including psychology, neuroscience, medicine, games and human-robot interaction [25, 96, 6]. In HRI, Several studies have explored the use of PBs in human-robot trust research, such as [89, 94, 117, 9, 115].

Khawaji et al. [94] investigated the potential of galvanic skin response (GSR) as a physiological measure of trust and cognitive load within a text-based chat environment. Participants, consisting of 28 university students paired into two groups, engaged in an investment game involving text-chat communication. Trust levels were manipulated by either allowing or restricting initial face-to-face interaction, while the cognitive load was adjusted by varying the complexity of the numerical tasks participants had to perform. The study design created four conditions: low trust-high cognitive load, low trust-low cognitive load, high trust-high cognitive load, and high trust-low cognitive load. Findings indicated that GSR responses aligned with the participants' levels of trust and cognitive load. Specifically, GSR values were lower in scenarios with high trust and low cognitive load, while higher values were associated with either high cognitive load or low trust. These results support the potential of GSR as an indicator of cognitive load and trust within overlapping conditions. However, the study's reliance on a single physiological metric within a controlled human-computer interaction environment and its limited sample size constrain broader

application, suggesting a need for future studies to validate findings across different communication modalities and physiological metrics.

Lu and Sarter [117] explored eye tracking as a measure of trust in automation. The study focused on how system reliability and priming influenced participants' trust calibration, using a UAV target detection simulation to observe these effects. Thirty-two participants monitored six UAV video feeds that presented targets or similar-looking distractors, and automation reliability varied between high (95%) and low (50%). Half of the participants received prior reliability information (priming) for certain UAVs, while others assumed reliability through direct observation. Various eye-tracking metrics, including fixation duration, saccades, and transitions, were analysed. Results indicated that participants' monitoring behaviour and visual attention aligned closely with system reliability. Low-reliability UAVs led to more frequent fixations, longer fixation durations, and increased transition rates, suggesting that participants trusted these UAVs less and thus monitored them more. Priming also influenced eye movements, with primed participants showing shorter saccades and reduced scan path lengths, implying a more focused monitoring strategy. Eye tracking data correlated well with subjective trust ratings, reinforcing eye tracking as a promising method for real-time trust calibration. However, the study's reliance on eye tracking without additional physiological indicators limits its comprehensiveness, and its applicability to dynamic real-world scenarios remains to be tested.

Gupta et al. [66] evaluated trust in a virtual assistant in VR using physiological measures. The research involved a dual-task environment where participants interacted with a virtual assistant during a VR shape-selection and N-Back task, simulating high and low cognitive loads. The virtual assistant's reliability was set to either high (100% accuracy) or low (50% accuracy), creating four conditions. Physiological data, including heart rate variability (HRV), GSR, and facial electromyography (EEG) signals, were collected to assess trust levels. Results indicated HRV as a reliable trust indicator, with elevated HRV correlating to higher trust in the virtual assistant. However, EEG and GSR measures did not consistently differentiate trust levels across conditions, suggesting the need for further study to refine physiological indicators of trust, especially in VR settings.

Hald, Rehmn, and Moeslund [68] designed a real-time, computer vision-based

system to measure human trust in a close proximity HRI setting. This assessment involved tracking participants' physical proximity and trust responses during a collaborative drawing task, where participants held paper as a robot arm drew between their hands. Trust was manipulated by unexpectedly increasing the robot's drawing speed midway through the task. GSR was used to measure emotional arousal, while an infrared camera tracked participant proximity to the robot. Results indicated that GSR was not significantly different across conditions, suggesting that GSR alone may be insufficient for assessing trust. Additionally, the motion tracking system, based on single-point tracking, did not produce significant results. The study highlights the need for a more comprehensive approach to physiological and behavioural trust measures, as well as tasks that allow for a broader range of physical reactions.

Zhang et al. [199] investigated dynamic trust changes in HRI, specifically comparing eye-tracking, electrocardiogram (ECG) metrics, and self-reported trust levels across phases of trust building, breach, and repair. In a collaborative assembly task with a Universal Robots UR10 robot, participants' trust was manipulated by changing the robot's reliability (100% vs 76%), and physiological responses were measured. Results showed that eye-tracking indicators, such as fixation counts and gaze transition entropy, captured trust dynamics effectively, particularly during trust breaches, while heart rate metrics from ECG lacked sensitivity to trust changes during assembly movements. The study highlights eye-tracking's potential to complement or even replace self-reports for tracking dynamic trust in HRI.

Machine learning to estimate Trust

The empirical studies discussed above provide strong evidence that various physiological behaviours, such as GSR, EEG, ECG, HR, and eye movements, show significant differences between states of trust and distrust during HRI. These findings motivate the use of machine learning (ML), which can develop more flexible and adaptive models capable of capturing the subtle and dynamic nature of human trust in robots over time.

Khalid et al. [89] developed a multi-dimensional approach to assess trust in human-robot-human interaction by integrating subjective measures of ability, benevolence, and integrity, which align with key aspects of truthfulness in

HRI, as defined by Mayer, Davis, and Schoorman [120], reflecting a robot's competence, ethical intentions, and reliability [120]. This framework was combined with objective physiological behaviours, including facial expressions, voice characteristics, heart rate variability (HRV), and skin conductance. In a controlled laboratory environment, 42 participants completed three tasks simulating various interaction scenarios, including video evaluations and face-to-face and teleoperated dialogues. The study utilised a neuro-fuzzy neural network model enhanced with evolutionary algorithms, achieving a 67% accuracy in trust classification. Their findings indicate that physiological cues, particularly facial expressions, vocal features, and HRV, can provide meaningful insights into trust levels within HRI. However, the model's limitations include the relatively low accuracy of 67%, potential inaccuracies in facial expression detection across diverse ethnicities, and the need for additional features to fully capture the complexity of trust dynamics in human-robot communication.

Hu et al. [82] developed a real-time trust sensor model that maps PBs to human trust levels during interactions with automated systems. In a simulated car-driving task, 31 participants evaluated the reliability of an obstacle-detection algorithm, choosing to trust or distrust the system based on its performance across reliable and faulty conditions. The study captured PBs using EEG and GSR to reflect arousal and attention, with specific features extracted from the frequency and time domains of these signals. Multiple classification methods, including support vector machines, logistic regression, and ensemble voting, were used to map EEG and GSR data to categorical trust levels. Ensemble voting demonstrated the best performance, achieving a mean classification accuracy of 71.57%. This demonstrates that PB measurements can be effectively used to sense trust in real-time.

Akash et al. [9] developed a real-time trust sensor using EEG and GSR to model human trust in human-machine interactions (HMI). In a simulated autonomous system, 45 participants responded to a sensor's accuracy in detecting obstacles, allowing for real-time classification of trust versus distrust states. The study compared a general model with an accuracy of 70% to a customised model that achieved 78.58% accuracy, though the customised model required additional training time. This work represents a pioneering effort to capture real-time psychophysiological responses to measure trust within HMIs, addressing the

situational and learned aspects of human trust.

Ajenaghughrure et al. [8] developed a predictive model to assess user trust in an AI (conversational user interface) using PBs. In their study, participants engaged in an information search game, where they answered questions with the help of Google Assistant. Trust dynamics were manipulated through question difficulty, the accuracy of the AI's responses, and risk factors associated with correct or incorrect answers. To capture trust levels, the authors collected EEG signals across eight electrodes, with feature extraction producing time and frequency domain metrics, including power spectra and statistical measures. After feature selection, four significant features were retained to train a voting classifier comprising seven algorithms. The model achieved a mean accuracy of 77.8%, demonstrating efficacy in identifying trust-related behaviours. Limitations included individual variability in trust responses and a small sample size (10 participants), with the authors suggesting future inclusion of additional physiological data, such as ECG or EDA, to improve stability and generalisability for real-time trust assessments.

Lochner, Duenser, and Sarker [115] investigated the calibration of human trust in automation using physiological indicators, specifically GSR, to measure trust and cognitive load in real-time during a semi-automated unmanned aerial vehicle (UAV) operation task. Forty-three participants navigated a UAV across points in a controlled indoor environment, rotating between high and low automation modes. Trust levels were manipulated through verbal and written cues to create high- and low-trust conditions. Using a decision tree model, the authors classified trust states with an accuracy of 80%. The study highlights the effectiveness of physiological data, especially GSR, in measuring trust and cognitive load but notes limitations in generalisability due to the experimental nature of induced trust and the structured UAV task.

Ajenaghughrure, Sousa, and Lamas [6] developed an ensemble trust classifier model to assess trust in real-time interactions within an AV simulation game. The study tested various psychophysiological signals, including EEG, EDA, ECG, facial EMG, and eye-tracking, to identify the most reliable signal for trust classification. Using a stacked ensemble model with a hybrid feature selection approach, the study achieved optimal performance with EEG (ranging from 60% to 80%) and multimodal EEG-based signals, suggesting EEG's high

temporal resolution is particularly effective for trust assessment. However, key limitations include the offline-only testing of models, which leaves their real-time performance uncertain, and the minimal inclusion of non-EEG features in the multimodal model, questioning the effectiveness of combining multiple signal types for trust assessment and highlighting the need for further exploration of multimodal feature selection strategies.

Xu et al. [187] aimed to recognise trust during human-robot cooperation using an EEG-based model within a cooperative game scenario designed to vary trust levels. Participants collaborated with robots in an Overcooked-AI game, providing a controlled, interactive environment for testing trust recognition. The study focused on EEG as the primary physiological measure of trust, utilising a Vision Transformer model with 3-D spatial representation to capture spatial EEG data. Results demonstrated high accuracy, with 74.9% in slice-wise validation and 62% in trial-wise validation, outperforming baseline models. However, limitations included individual performance variability and reduced generalisability for real-world, cross-trial applications.

Section 2.3.2 Gap:

In summary, the described empirical studies have employed various PBs, such as GSR, HRV eye fixation, and EEG, to assess trust in collaboration contexts only in a single interaction. Besides, existing work on the use of machine learning to estimate trust has used limited PBs in combination and has also created a dataset based on one-off interaction. In addition, the majority of these studies had simulated environments rather than real HRI. Moreover, existing research has mainly focused on collaborative settings, with the only exception we identified being [141], which investigates trust in competitive contexts. In competitive contexts, factors influencing human trust, such as the robot's truthfulness in providing advice or information and the capability of robots to outperform humans, may not be present in collaborative settings. Moreover, reported accuracy rates in these studies range from a low of 60% to a high of 80%, 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. Furthermore, the limited number of participants, low as 10 to a maximum of 45 in these studies, restricts the generalisability of findings to border scenarios. Another limitation is that these studies were conducted in one-off interactions, which do not account for the dynamic of trust over repeated interaction.

This thesis seeks to bridge these gaps by developing a trust assessment framework that encompasses both collaborative and competitive HRI contexts, incorporating PBs from a broader range of sources and tracking trust dynamics over repeated interactions. This research provides a robust model with enhanced adaptability by utilising two diverse datasets from repeated collaborative and competitive tasks and employing incremental transfer learning to improve prediction accuracy over time. With a more comprehensive participant pool of 85 individuals, this work aims to refine trust measurement in HRI, offering an adaptable and dependable model suited for broader real-world applications.

2.3.3 Trust And Vocal and Non-vocal cues

Vocal and non-vocal cues used in predicting trust appear in various domains, where facial expressions and vocal characteristics correlate with human behaviours of trust and distrust. For instance, facial expressions and blend shapes, often provide subconscious signals of a person's emotional state and level of trust or discomfort in a given situation [101]. Positive expressions, such as smiling, are frequently linked to higher levels of trust [166]. Similarly, vocal cues, including tone of voice and speech patterns, play a crucial role in trust dynamics [175]. For example, higher pitch and slower speech rates have been identified as indicators of trust in specific scenarios, such as economic games [176]. These cues can be automatically detected through cameras with a microphone, making them useful for real-time trust assessment. Considering these empirical findings, few studies in HRI have attempted to explore the use of vocal and non-vocal cues to measure trust in HRI. While these studies showed promising findings, to our knowledge, only two studies explored the use of facial expressions [89, 30], two explored vocal cues in measuring trust [49, 60], and only one study combined vocal and non-vocal cues to assess trust in HRI [89].

Elkins and Derrick [49] explored the relationship between vocal dynamics and trust during interactions with Embodied Conversational Agents (ECAs) by examining how vocal cues, such as pitch and response duration, and ECA features, such as gender and smiling, impact perceived trust. The study was conducted in a controlled laboratory environment, and participants engaged in a simulated screening interview where the ECA's behaviour and gender were manipulated randomly across interview blocks. The study found that trust in the ECA increased over time, with smiling ECAs perceived as more trustworthy than neutral ones. Additionally, higher vocal pitch was initially associated with lower trust levels. However, this effect lessened as interactions continued, indicating that trust-building is dynamic. While this study reveals the significance of temporal and expressive factors in trust development with ECAs, it has several limitations. The reliance on basic vocal measures and one-time interactions may not fully capture trust dynamics in extended or repeated engagements. Furthermore, with a limited range of ECA characteristics manipulated, findings may not generalise to more diverse virtual agents or naturalistic settings. These results underscore the need for additional cues and

repeated or prolonged approaches to understanding better trust in human-agent interactions, particularly across varied cultural and situational contexts.

Gauder et al. [60] investigated trust detection in human-virtual assistant interactions, focusing on how a user's trust in a VA's competence could be inferred from vocal cues in their speech. The study deployed a novel protocol where 84 participants (both in-lab and remote) interacted with VAs labelled as either "competent" or "incompetent." Participants' speech was recorded as they asked and responded to factual questions answered by the VA, designed to elicit either trust or distrust in the VA's abilities. Vocal cues such as speech rate, pitch, pause duration, and hyper-articulation were measured to detect perceived trust levels. Using a classification model based on these vocal features, the study achieved an accuracy of up to 76% in identifying whether participants believed the VA to be competent. However, limitations included the variability in recording conditions for remote participants and the narrow focus on competence as a representative for trust, which may not fully capture the complexity of trust in automation. Overall, the study highlights the promising potential for using vocal cues to measure trust in automated systems, suggesting applications in adaptive trust-based HRI design.

Campagna, Chrysostomou, and Rehm [30] explore the relationship between facial expressions and trust levels in an industrial human-robot collaboration setting, aiming to improve trust assessment accuracy in collaborative robotics. They conducted the study in a controlled chemical industry simulation involving a Universal Robots UR10-CB3 collaborative robot assisting participants in chemical handling tasks. The robot's performance varied between high and low-trust conditions, with trust categorised based on participant reactions. Using Convolutional Neural Networks (CNNs) to analyse facial expressions, the study achieved an accuracy of 78.61% in identifying trust during the handling task and 73.35% during the pouring task, highlighting the effectiveness of deep learning methods for facial expression-based trust evaluation. However, this study was conducted in one-off interaction with a small sample size of 20 participants, which may limit the generalisability of these findings. Moreover, the study suggests incorporating sensor fusion for enhanced trust assessment robustness, emphasising the need for multimodal cues beyond facial features alone to address the complex, dynamic nature of trust in real-time HRC scenarios.

Section 2.3.3 Gap:

These studies generally focus on a limited set of either vocal *or* non-vocal cues, often neglecting the integration of multiple indicators that could provide a richer understanding of trust. Furthermore, trust is typically measured in single, one-off interactions, which overlooks its evolution over time and limits the generalisability of the findings. In this thesis, we address these gaps by integrating a wide range of vocal and non-vocal cues, including seven different facial expressions (e.g., happiness, sadness, and surprise), 52 facial blend shape features (capturing muscle movements like eyebrow raises and lip movements), and 13 specific speech characteristics (such as pitch, duration, and spectral bandwidth). By incorporating this diverse set of indicators, we aim to deliver a more comprehensive and dynamic model for predicting human trust in robots, capable of capturing its progression across repeated HRI interactions.

2.4 Trust in Repeated or Long-term interactions

Modelling and measuring human trust in repeated or long-term interactions allows for a structured evaluation of trust dynamics over time, which is essential for understanding how human trust in robots evolves and adapts based on experience.

2.4.1 Trust Framework

Hoff and Bashir [78] presented a framework for conceptualising trust that has been widely reported in the HRI literature [163, 155, 125]. 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 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 [154, 43]. The mental well-being of a human is also essential. Humans in a pleasant mood are likelier to have an initial trust in robots [169]. **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 [71]. Experience in robots can be built based on robot performance in previous interactions with robots in a particular context [158].

2.4.2 Empirical Studies on Trust in Repeated or Long-term Interactions

Building on the multi-layered trust framework, limited empirical studies provide insights into how dispositional, situational, and learned trust evolve during repeated interactions, highlighting both the potential and limitations of existing approaches in capturing the complexities of long-term HRI.

Miller et al. [125] studied how different layers of trust—dispositional trust, initial learned trust, and dynamic learned trust—interact and influence user trust in robots during initial interactions. This research was conducted in a laboratory setting using a Wizard of Oz paradigm, where participants interacted with a domestic humanoid robot named TIAGo, which was programmed to approach them twice. During each trial, participants reported their trust levels and indicated a comfortable stopping distance as the robot approached. The study aimed to discover the factors influencing trust in robots, focusing on the roles of user disposition, anxiety, and attitudes towards robots. Findings revealed that dispositional trust in automation and prior negative attitudes toward robots significantly influenced both initial and dynamic learned trust, where anxiety also played a critical role. Initial learned trust acted as a baseline that informed dynamic trust through the interactions, with participants allowing the robot closer over time, suggesting a habituation effect. Notably, dynamic learned trust increased with each interaction, suggesting a growing familiarity and

comfort. Although these insights are valuable, the sample size of 28 participants is relatively small, and the short-term nature of the interaction is a limitation.

Yanco et al. [192] conducted a study to investigate whether an operator's control allocation strategy and trust level change over an extended period. The research found that there were no significant differences in any attribute between sessions two through six, indicating that an operator's behaviour during initial interaction can predict their behaviour over the short term. However, the study observed a notable difference in operators who were familiar with robots; they displayed less trust and experienced an increased workload. Although the model provides valuable insights into trust dynamics in HRI, it cannot be used dynamically to predict the current levels of user trust during robot use.

Hafizoğlu and Sen [67] examined how prior experiences shape trust in repeated human-agent teamwork within virtual environments. Their study involved participants interacting with trustworthy or untrustworthy agent teammates in a repeated coordination game called the Game of Trust (GoT). Participants worked alongside an agent teammate without prior communication or subtask assignments, fostering an environment where initial trust and adaptability were crucial. The results revealed that participants with positive past experiences showed increased trust in agent teammates, whereas those with negative experiences exhibited decreased trust. This change in trust, influenced by emotional state, task expertise, and participant expectations, was consistent across multiple sessions, though its intensity diminished over time. While these findings underscore the importance of designing agents that adapt to humans' trust-related behaviours, the study's reliance on virtual settings without physical embodiment or direct communication limits the direct applicability of findings to real-world, physically co-present human-agent interactions.

Gremillion et al. [63] proposed a recursive estimation framework that can predict changes in driver authority in human-autonomy driving interactions, specifically focusing on toggling authority with an autonomous driving assistant. Using a simulated two-lane closed circuit, participants decided when to engage or disengage the driving assistant, while physiological data, including EDA, HR, and EEG, were captured alongside environmental inputs such as lane positioning and traffic conditions. The study employed stochastic filtering methods, notably particle and unscented Kalman filters, to recursively estimate

the driver's trust-based decision state. Their model demonstrated an ability to predict drivers' decisions to toggle authority based on physiological and contextual cues, marking a significant advancement in interpreting real-time trust indicators within autonomous systems. However, the model's accuracy and applicability were constrained by the need for self-reports and its confinement to a simulated environment, which may limit generalisability to real-world settings.

Rossi et al. [149] examined how the timing of errors by a companion robot influences human trust in long-term interactions. Using the Care-O-bot 4 in a simulated home environment, six participants interacted with the robot across multiple sessions over three weeks. The study implemented two conditions: errors with severe consequences (e.g., revealing private information or failing to switch off gas) occurred either at the beginning or the end of the sessions. In the concluding session, participants faced a simulated emergency (a fire) to assess their trust in the robot's capability to respond. Results indicated that trust was more negatively affected when errors occurred at the start, with participants showing less trust recovery in later interactions. Conversely, end-session errors had a milder effect on trust. These findings underscore the importance of initial interactions in establishing trust but are constrained by the study's limited sample size and specific error types, suggesting a need for further research to confirm these dynamics across broader contexts and populations.

Similarly, we see limited work on studying and validating trust models during repeated interactions [42, 186, 35, 40].

Desai [42] was the first to model trust in repeated HRI with autonomous robots. In their findings, trust ratings remained relatively consistent over time, suggesting that users' trust levels, once established, did not fluctuate significantly across interactions. Furthermore, familiarity with the robot emerged as a critical factor influencing trust: participants who were more familiar with the robot tended to trust it less over time compared to those with limited exposure. This pattern suggests that greater familiarity may lead to a heightened awareness of the robot's limitations, thereby fostering a more cautious trust calibration. Desai's work highlights the importance of incorporating familiarity and repeated exposure when modelling trust for long-term HRI, as these factors can affect trust consistency and user reliance on the system.

Chen et al. [35] aimed to enhance long-term human-robot collaboration by incorporating trust as a dynamic variable in the robot's decision-making process. The study introduced a trust-based reinforcement learning model that enabled the robot to adjust its actions based on the anticipated level of human trust during repeated interactions. In a real-world table-clearing task involving 20 participants, individuals decided when to allow the robot to perform specific actions or when to intervene. The results indicated that the trust-RL model improved collaboration over time, reducing unnecessary interventions and enhancing overall team efficiency. However, the model's limitations included a simplified representation of trust and an assumption of the robot's success, which may not fully reflect the complexities of trust dynamics in more intricate real-world situations involving possible robot failures.

More recently, De Visser et al. [40] proposed a model for trust calibration within long-term human-robot teams, integrating concepts such as relationship equity, social collaborative processes, and self-perception. This model provides a systematic approach for trust to evolve naturally over repeated interactions, considering moment-to-moment adjustments based on mutual perceptions and expectations. A unique feature of this model is its emphasis on relationship equity, an emotional "account" that can predict friendship over time and helps manage both overtrust and undertrust within HRI. The model was evaluated over ten interaction rounds, focusing on real-time adjustments based on observed trust calibration needs. However, certain limitations were noted in the study: interactions were primarily simulated with a small sample size, and the robot's representation was image-based rather than physical, which may not fully capture the complexities of physical HRI.

Section 2.4 Gap:

In summary, while prior research has provided foundational insights into trust dynamics in repeated or long-term HRI, significant limitations remain. Studies like those by Miller et al. [125] and Yanco et al. [192] underscore the roles of dispositional trust, familiarity, and initial interactions in shaping trust. However, short-term laboratory settings, small sample sizes, and simplified tasks often constrain these studies, limiting validity. Additionally, existing trust models, such as those by Desai [42] and Chen et al. [35], often rely on fixed initial trust values and static assumptions, failing to account for the evolving nature of trust across different contexts and repeated, real-world interactions. Key factors, such as risk, ambiguity aversion, and real-time adaptability, also remain underexplored in many of these models. This thesis aims to address these gaps by developing a mathematical trust model that dynamically incorporates user experience based on robot performance, user control, risk, and ambiguity aversion over repeated interactions with physical robots. By validating this model in both collaborative and competitive contexts, we aim to capture a more comprehensive view of trust dynamics over time. In addition, we employ physiological behaviour measurements (e.g., EEG, GSR, HR) and monitor vocal and non-vocal cues (such as facial expressions and voice intonations) across these repeated interactions to provide a richer, real-time understanding of trust changes. This integrated approach not only takes into account a wider range of trust-related factors but also utilises various indicators that capture subconscious and emotional responses to HRI. By merging these methods, our model aims to provide a strong and ecologically valid framework that better reflects the complexities of trust in repeated and long-term HRI.

2.5 Reinforcement Learning For Optimising Trust in HRI

Reinforcement learning (RL) is a machine learning technique that trains software agents to make optimal decisions in order to achieve the maximum rewards within a specific environment [172]. RL algorithms are inspired by behavioural

psychology, mirroring how humans and animals learn through trial and error [172]. In RL, the agent explores an environment by taking action, receiving feedback (rewards or penalties), and updating its policies to improve future decision-making [111]. This trial-and-error approach allows agents to discover optimal or near-optimal strategies in complex, dynamic environments where explicit programming of all possible scenarios is not feasible [196]. This capability makes RL particularly well-suited for applications where dynamic and adaptive behaviour is required—such as in HRI, where robots need to gain trust and act reliably based on human feedback. RL methods have been widely adopted to enhance human-robot collaboration, particularly in autonomous systems and decision-making scenarios [197, 196]. Traditionally, RL methods optimise robot behaviours to maximise reward functions tied to performance metrics, such as task completion speed or resource efficiency. Some studies have explored human-centred approaches to RL, where human feedback is explicitly incorporated into the learning process. For instance, Knox and Stone [97] introduced the concept of Interactive RL, where human feedback is used to guide the robot’s learning process. In multi-agent systems, RL has been utilised to foster interactions between humans and autonomous agents, with trust emerging as a crucial element of effective collaboration. For instance, Omidshafiei et al. [134] implemented deep reinforcement learning in multi-agent settings where the trustworthiness of agents influenced their cooperative strategies. Although their focus was primarily on agents rather than human trust, such studies provide valuable insights into how RL can navigate complex trust dynamics in collaborative environments. Moreover, in collaborative environments, such as manufacturing and assembly, RL has been employed to enable robots to predict human intent and adjust their actions accordingly. For instance, Liu et al. [113] introduced an RL-based framework that allows robots to learn how to anticipate and complement human actions in a shared workspace.

Section 2.5 Gap:

Although RL has indirectly impacted trust by improving overall collaboration, it has not traditionally been utilised to dynamically adjust trust during interaction. One notable exception is the work by Abouelyazid [2], which presents a reinforcement learning framework designed to enhance both safety and trust in HRI. This framework incorporates constrained reinforcement learning to address safety concerns alongside reward shaping to include trust-building behaviours directly. The study also highlights the importance of interactive feedback through human-in-the-loop learning, enabling robots to adapt their actions dynamically based on human responses. However, as this framework remains theoretical, it lacks empirical validation, raising questions regarding its applicability in real-world scenarios.

These gaps highlight the need for a more comprehensive framework that optimises trust calibration. This ensures that human trust in robots aligns appropriately with the robots' actual capabilities and limitations, thereby reducing the risks associated with over-trust and under-trust.

This work uses a validated mathematical trust model combined with RL to optimise human trust during HRI. Unlike existing RL methods, which primarily enhance performance, our approach directly targets dynamic trust calibration, ensuring trust is well-balanced for improved collaboration. This provides a novel and effective method to manage trust in HRI, offering practical applications for real-world human-robot teamwork.

Chapter 3

Factors Affecting Human Trust in Robots

This chapter ¹ presents the first empirical investigation in this thesis, focusing on understanding the factors that influence human trust in robots and how these factors vary across different contexts and cultures to effectively model human trust in robots. Although existing models, such as those proposed by Hancock et al. [70, 71], offer a solid foundation for understanding trust in human-robot interaction (HRI), recent research suggests that these factors can vary significantly based on the specific HRI scenario and the level of risk involved [1, 171]. Additionally, the impact of cultural differences on trust remains an underexplored area, particularly in non-Western contexts such as the Arab world.

In this chapter, we present two studies conducted in Saudi Arabia (KSA) and the United Kingdom (UK). We selected these two countries because limited studies can be found where Arab culture is compared to European/Western culture in the context of studying trust in robots [71]. Furthermore, the use of robotics is among the top government strategies in different domains, such as manufacturing, defence, and nuclear power, for the near future of the UK [114] and KSA [74]. These studies investigate how trust in robots varies across three different HRI scenarios: guiding visually impaired individuals, teleoperated healthcare diagnostics, and industrial assembly tasks. Each of these scenarios

¹This Chapter has been published as a conference paper:

- Alzahrani, A., Robinson, S., & Ahmad, M. (2022, December). 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).

differs in the degree of risk, task complexity, and interaction type, allowing a detailed analysis of trust factors in diverse environments. Additionally, the studies compare how trust is perceived in different cultural contexts.

This chapter addresses the following research questions:

- RQ1: Do individuals' trust perception of a robot vary across HRI settings?
- RQ2: Do people from different cultural backgrounds show differences in factors affecting trust, and does their perception of trust vary across diverse settings?
- RQ3: What new or previously unidentified factors influence user trust in robots across these settings?

These research questions provide essential insights for the development of the mathematical trust model in Chapter 4. Understanding the factors that influence trust in different HRI scenarios (RQ1) ensures that the model can adapt dynamically to context-specific variations in trust perception. Identifying cultural differences in trust factors (RQ2) enables the model to be more generalisable, accounting for diverse user expectations and avoiding cultural bias in trust calibration. Additionally, uncovering new or previously unidentified trust factors (RQ3) refines the inputs for the trust model, ensuring it is comprehensive and accounts for real-world influences that existing models may have overlooked. These findings collectively contribute to the formulation of a robust mathematical model that effectively predicts and adjusts trust levels in human-robot interactions across varying conditions.

The novel contributions of this chapter are:

- We show that not all factors affecting trust listed in the Hancock et al. [71] model are relevant or essential across various scenarios and cultures.
- We show that factors affecting trust (such as controllability, familiarity, usability and others (see Table 3.3 & Figure 3.4)) vary across scenarios in both studies and also varied across two different cultures (KSA and the UK).

- We show that the trust perception and trust relevance scores varied significantly across the two different cultures. However, trust perception did not vary across scenarios in the study conducted in the UK.
- We identified new factors (controllability, familiarity, and risk) in addition to ones identified by [71] in both studies.

3.1 Study Design

We conducted two parallel online studies using a mixed-methods research design to investigate how trust in robots varies across different HRI settings and cultural contexts. Study one was conducted in Saudi Arabia (KSA) and study two in the United Kingdom (UK), allowing for cross-cultural comparison. The research employed a structured comparative approach examining three distinct human-robot interaction scenarios (guiding visually impaired individuals, teleoperated healthcare diagnostics, and industrial assembly tasks), each representing different levels of risk, task complexity, and interaction types.

Each study followed an identical three-phase workshop format:

1. **Introduction & Consent:** Participants were introduced to the study, provided with an overview of the research objectives, and signed consent forms.
2. **Scenario Evaluation:** Participants watched three videos depicting different HRI scenarios (guiding visually impaired individuals, teleoperated healthcare diagnostics, and industrial assembly tasks). They then completed trust-related questionnaires to assess their trust perception (TPS) and trust relevance (TRS) for each scenario.
3. **Focus Group Discussion:** A structured discussion was conducted where participants reflected on the factors influencing their trust in robots across the three scenarios. This session provided qualitative insights into human, robot, and environment-related trust factors.

This design allowed for both quantitative analysis, through the collection of TPS and TRS scores, and qualitative analysis, through thematic coding of participants' discussions. The results were used to test the following hypotheses:

We conducted two online studies. Study one was conducted in KSA, and study two was conducted in the UK. These studies aimed to investigate how trust in robots varies across different HRI settings and cultural contexts. Each study consisted of a structured workshop format, which included three phases:

H1: The number of factors affecting participants' trust will differ across human, robot, and environment-related factors in the three HRI settings.

H2: Participants' trust perception score (TPS) (**H2a**) and trust relevance score (TRS) (**H2b**) in the robots will vary significantly across the three HRI settings.

H3: Participants' trust perception score (TPS) and trust relevance score (TRS) will vary significantly across cultures.

3.1.1 Ethics

Two applications were submitted to the university ethics board to ensure ethical integrity. Following a review process, they were approved [160322/4758]. Participants were provided with a participant information sheet outlining the study objectives, procedures, and data confidentiality measures. Prior to participation, they gave informed consent by signing a digital consent form, which explicitly stated their right to withdraw at any time without consequences. The study adhered to institutional ethical guidelines for research involving human participants.

3.1.2 Task

The task followed twelve different steps as illustrated in Figure 3.1 and the task presentation that we used.² The task involved watching three videos of human-robot interaction in three different scenarios.

1. Guiding: a dog guide robot led a blind person through narrow and clustered spaces to his destination [183].
2. Healthcare: a teleoperated robot that assists medical practitioners in diagnosing patients remotely with a nurse's assistance [185].

²https://docs.google.com/presentation/d/1UIB3QNgHHkkgqCGz1s_K7Dm3q3yLTD0m/preview

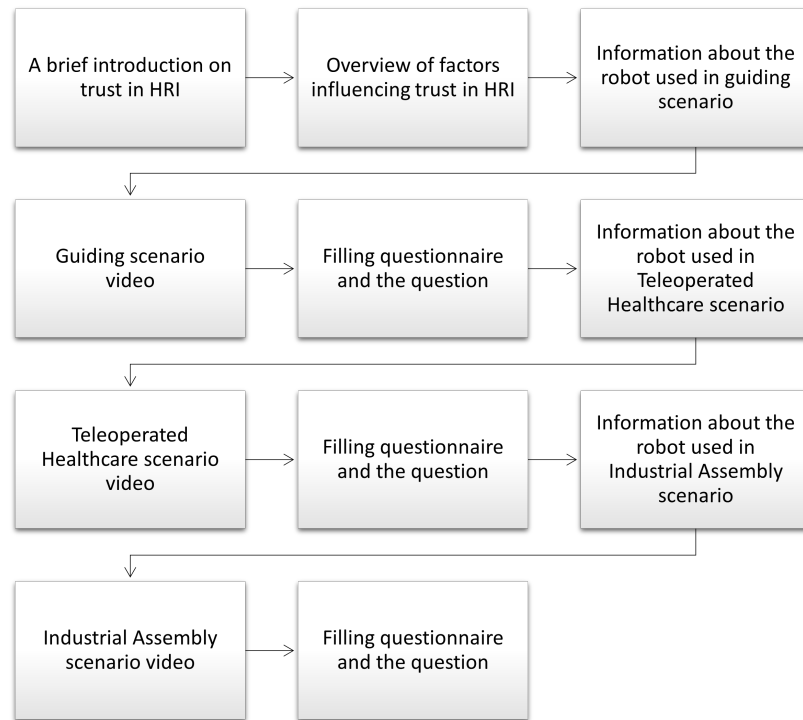


FIGURE 3.1: The steps taken in the workshop (upper left to lower right).

3. Manufacturing: a manufacturing robot collaborates with workers on an assembly line to construct a product [45].

Before engaging with the tasks, participants were provided with a definition of "robot" to ensure a shared understanding of the term, which describes a robot as a mechanised system, typically programmable via a computer, that can execute a sequence of complex actions. [104].

The tasks were always presented in this order for all participants to ensure consistency in task delivery.

3.1.3 Participants

In study one, we recruited 18 participants (Mean age: 35.16 years, SD = 6.88, 44% female) from KSA. 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 a few times. 3 participants had high experience interacting with robots. 4 participants had medium experience interacting with robots. 8 participants had low experience interacting with robots. 6 participants had no experience interacting with robots. In study two, we recruited 18 participants (Mean age: 27.77 years, SD = 7.21, 56% female) from the UK. 4 participants had high experience interacting with robots. 5 participants had medium experience interacting with robots. 4 participants had low experience interacting with robots. 5 participants had no experience interacting with robots. It is worth noting that the experience with robots was similar among participants in both studies. All the participants were postgraduate students and academics in the computer science department at their respective universities. This selection was intentional, as familiarity with robotics was essential for participants to critically evaluate and articulate the factors influencing human trust in robots. By selecting individuals with prior knowledge in robotics and technology, we ensured that responses were informed and contextually relevant. Participants were invited via email and flyers. In each study, we conducted the workshops with a group consisting of three participants at a time. The registration for the study was managed using an online application for registration (*Calendly*³). The workshop was conducted online via Zoom with all the participants. During these Zoom sessions, participants individually completed all questionnaires using Google Forms, which allowed for private and independent responses without influence from other participants. The researchers provided instructions for each questionnaire through screen sharing, and participants were given sufficient time to complete them at their own pace. After watching each scenario video, participants individually filled out the TPS and TRS questionnaires before proceeding to the next scenario. This individual completion approach ensured that each participant's trust perceptions were captured independently before the group discussions began. The completed questionnaire responses were automatically recorded in a secure database for subsequent analysis. This approach maintained data integrity while facilitating the remote nature of the study during pandemic restrictions.

³<https://calendly.com>

3.1.4 Procedure

The study was conducted in two parts: 1) evaluating in the three HRI scenarios; and, 2) a focus group activity. Participants initially received a participant information sheet, consent form, and Zoom link before each meeting. In the evaluating HRI scenarios, participants completed the following steps:

1. Participants completed the demographics questionnaire.
2. Participants watched the HRI video.
3. Participants completed the questionnaire to rate the robot using TPS and TRS, respectively.
4. Participants wrote the factors affecting their trust in robots in the demonstrated interaction.
5. Participants repeated steps 2, 3, and 4 for the other two scenarios.

Focus group activity: At the end of the workshop, we used the mini-group discussion method. The three participants with sufficient knowledge of the topic were asked to discuss the factors affecting trust in the three different scenarios after watching each video in the light of Hancock's model of trust. The groups discussed the following themes in each scenario: 1) Human-related factors, 2) Robot-related factors, and 3) Environment-related factors. The participants had an opportunity in each scenario to discuss their thoughts, and we recorded the discussion to analyse the data. The average duration of each group discussion was 20 minutes.

3.1.5 Measurements

To measure trust and examine changes in trust between different situations, we used the TPS developed by Schaefer [158]. We asked participants to complete the TPS questionnaire to rate their trust in the robot in each scenario. The scale has 40 items and a subscale of 14 items (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 as a percentage. This study used the 14-item subscale because

it had the most relevant factors according to the scenarios depicted in the video. Following the instruction given in [158], we computed the trust score by first reverse coding the 'have errors,' 'unresponsive,' and 'malfunction' items, then summing the 14 item scores and dividing by 14.

To measure the relevance of trust, we asked participants to rate the relevance of trust in each scenario on a 5-point Likert-like scale, ranging from *not at all relevant* to *very relevant*. We also analysed the factors influencing trust by categorizing them into three themes: human-, robot-, and environment-related factors. These factors were coded based on their frequency across the three settings.

For the qualitative analysis, we used NVivo 12 software to systematically code and explore participants' responses regarding the factors influencing human trust in robots. The primary goal of the qualitative analysis was to identify patterns, uncover new or previously unconsidered trust factors, and understand the reasoning behind participants' trust perceptions in different scenarios. This qualitative approach provided contextual insights that complemented the quantitative findings, ensuring a more comprehensive understanding of trust dynamics across diverse HRI settings.

3.2 Results

This section presents both quantitative and qualitative findings from our study, offering a comprehensive view of the factors shaping trust in HRI across different scenarios and cultural backgrounds. The quantitative findings examine statistical variations in trust perception (TPS) and trust relevance (TRS) scores, highlighting key differences across settings and cultures. The qualitative findings provide deeper insights by analysing participants' perspectives, categorising trust factors into human-, robot-, and environment-related themes, and identifying emerging influences on trust perceptions.

3.2.1 Quantitative findings

Study 1 (Saudi Arabia)

To test **H1**, we conducted a Chi-Square Goodness of Fit Test to determine whether the frequency of human, robot and environment-related factors was

Scenario	Study 1			Study 2		
	Human	Robot	Environment	Human	Robot	Environment
1	3	22	4	4	33	5
2	5	15	1	7	28	9
3	2	14	2	5	25	7

TABLE 3.1: Frequency of factors affecting trust across the three scenarios in study 1(KSA) and study 2(UK).

equal between the three scenarios. The Chi-Square Goodness of Fit Test is appropriate for this analysis as it assesses whether the observed distribution of categorical variables (in this case, the frequency of identified trust factors) significantly deviates from an expected uniform distribution. The frequency data can be seen in Table 3.1. We set the significance threshold (α) at 0.05, as this is a commonly accepted standard in HRI research [78, 160] and has been consistently applied throughout this thesis. We did not find a significant difference among the frequencies of human-related factors, robot-related and environment-related factors across the three scenarios: $\chi^2(2, 18) = 1.40, p = 0.49$. $\chi^2(2, 18) = 2.23, p = 0.33$, $\chi^2(2, 18) = 2.0, p = 0.37$.

Before conducting parametric tests such as ANOVA and t-tests, we assessed the normality of our data using the Shapiro-Wilk test and visual inspection (histograms and Q-Q plots). The results confirmed that the data followed a normal distribution, as the p-values from the Shapiro-Wilk test were non-significant ($p > 0.05$), indicating that we failed to reject the null hypothesis of normality. Distribution plots of the Trust Perception Score (TPS) and Trust Relevance Score (TRS) are shown in Figures 3.2 and 3.3.

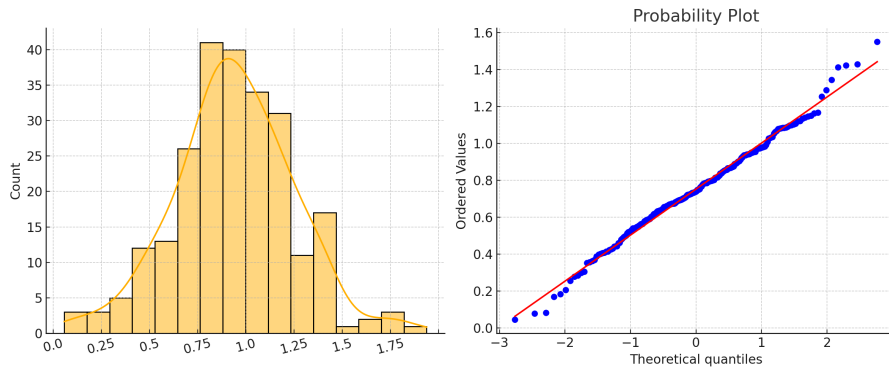


FIGURE 3.2: Histograms and Q-Q plots for TPS

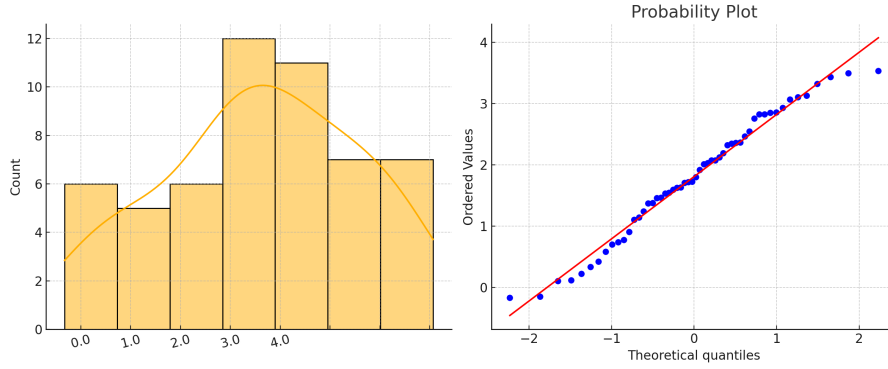


FIGURE 3.3: Histograms and Q-Q plots for TPS

Scenario	N	TPS score M		TPS score SD		TRS score M		TRS score SD	
		Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
1	18	0.76*	0.78*	0.15	0.11	4.11*	4.67*	0.96	0.69
2	18	0.66**	0.77**	0.16	0.13	3.50*	4.11*	1.04	1.28
3	18	0.78*	0.82*	0.12	0.17	4.83***	3.72***	0.38	1.07

TABLE 3.2: Mean (M) and standard deviation (SD) of TPS score and relevance score across the three scenarios in Studies 1 and 2. Bold values indicate significantly higher scores in cross-cultural comparisons. Significance levels: * $p \leq 0.05$, ** $p \leq 0.03$, *** $p \leq 0.001$.

To test **H2**, we conducted a one-way analysis of variance (ANOVA) to compare the TPS and TRS across the three scenarios. ANOVA is suitable in this context because it determines whether there are statistically significant differences between the means of three or more independent groups (i.e., the three different HRI scenarios). Since TPS and TRS are continuous variables, ANOVA allows us to assess whether scenario type has a meaningful effect on trust ratings. We found a statistically significant effect of scenarios on the TPS across the three conditions $F(2,51) = 3.9$, $p < 0.05$. We also conducted a post-hoc test to understand the difference between scenarios. The test indicated that the TPS for Scenario 2 was significantly lower than for Scenario 1 and Scenario 3. However, the TPS in scenario 1 did not differ significantly from scenario 3.

We found a statistically significant effect of scenarios for TRS across the three conditions $F(2,51) = 11.11$, $p < 0.05$. A post-hoc analysis indicated that the trust relevance for Scenario 3 was significantly higher than for Scenario 1 and Scenario 2. However, the trust relevance in scenario 1 did not differ significantly from scenario 3. The mean (M) and standard deviation (SD) can be found in Table 3.2.

Study 2 (United Kingdom)

To test **H1**, we conducted a Chi-Square Goodness of Fit Test to determine whether the frequency of human, robot, and environment-related factors was equal between the three scenarios. The frequency data can be seen in Table 3.1. We did not find significant differences among the frequencies of human-related factors, robot-related and environment-related factors across the three scenarios, $\chi^2(2, 18) = 0.8, p = 0.64$. $\chi^2(2, 18) = 1.14, p = 0.56$, $\chi^2(2, 18) = 1.14, p = 0.57$.

To test **H2**, we conducted a one-way analysis of variance (ANOVA) to compare the TPS and trust relevance score across the three scenarios. We found a non-significant effect of scenarios on the TPS across the three conditions $F(2,51) = 2.670, p = .079$. We found a statistically significant effect of scenarios for trust relevance across the three conditions $F(2,51) = 3.74, p < 0.05$. A post-hoc analysis indicated that the trust relevance score for Scenario 1 was significantly higher than that of Scenario 3. The M and SD can be found in Table 3.2.

Results – comparing Studies 1 and 2

To test **H3**, an independent samples t-test was conducted to compare the TPS and trust relevance score across the two studies. The independent samples t-test is appropriate because it compares the means of two independent groups to determine whether there is a statistically significant difference between them.

The findings show there was a significant difference in the TPS for all three scenarios, with UK participants (Study 2) consistently showing higher trust perception than KSA participants (Study 1): scenario 1 ($t(34) = -2.0, p = 0.05$, UK median = 0.79, KSA median = 0.74), scenario 2 ($t(34) = -2.23, p = 0.03$, UK median = 0.78, KSA median = 0.65), and scenario 3 ($t(34) = -2.08, p = 0.04$, UK median = 0.83, KSA median = 0.77).

For trust relevance scores (TRS), there were also significant differences between the two studies, with UK participants rating trust as more relevant in scenarios 1 and 2, while KSA participants rated trust as more relevant in scenario 3: scenario 1 ($t(34) = -1.9, p = 0.05$, UK median = 5.0, KSA median = 4.0), scenario 2 ($t(34) = -1.87, p = 0.05$, UK median = 4.5, KSA median = 3.5), and scenario 3 ($t(34) = 4.13, p < 0.001$, KSA median = 5.0, UK median = 4.0).

	Factors	Study1	Study2	%
Human	Controllability	5	5	59%
	Familiarity	1	4	29%
	User's condition	0	1	6%
	Accountability	0	1	6%
Robot	Robot's energy source	6	2	36%
	Consistency	-	7	32%
	Usability	-	4	18%
	Noise	1	1	9%
	Brand value	1	-	5%
Environment	Risk	1	8	75%
	Choice of use	-	2	17%
	Clarity of task	-	1	8%

TABLE 3.3: Frequency of new factors affecting trust across the three scenarios in study 1 and study 2 identified by the participants.

Overall, both TPS and TRS vary significantly across cultures, with UK participants generally showing higher trust perception scores across all scenarios, while trust relevance varied by scenario type. The mean, standard deviation, and significance indicators for both scores in both studies can be seen in Table 3.2.

3.2.2 Qualitative findings

The analysis process was as follows:

1. We transcribed the audio of the group discussion and uploaded transcription files to NVivo.
2. We created the themes that were derived from the Hancock model: human-related, robot-related, and environment-related factors.
3. Author 1 coded the transcripts.
4. Author 3 reviewed the codes to ensure they were relevant to the trust factors and assigned them to the appropriate themes. Here we discuss both studies' results for each theme and later compare with each other. The participants for both studies were coded as Participant (P), P1, P2, P3, ..., P18, respectively.

We have made the data analysis sheet available to ensure transparency and an open science framework.⁴ In Table 3.4, based on the qualitative analysis, we list participants' identified factors influencing human trust in robots.

Themes	Sub-themes	Study 1 - Initial codes (Frequency)	Study 2 - Initial codes (Frequency)
Human-related factors	Ability-Based	Controllability (5)	Controllability (5)
		Prior experiences (2)	Prior experience (1)
		Situation awareness (2)	Situation awareness (2)
		Familiarity (1)	Familiarity (4)
			Accountability (1)
			User's condition (1)
Robot-related factors	Characteristics-Based	Culture (2)	Culture (2)
	Performance-Based	Behaviour (6)	Behaviour (20)
		Reliability (6)	Reliability (20)
		Predictability (2)	Predictability (1)
		Mode of communication (2)	Mode of communication (8)
		Failure rate (19)	Failure rate (18)
		Noise (1)	Noise (1)
		Robot's energy source (6)	Robot's energy source (2)
		Level of automation (1)	Dependability (3)
			Consistency (7)
			Usability (4)
			Adaptability (1)
	Attribute-Based	Adaptability (4)	Adaptability (1)
		Robot type (1)	
		Brand value (1)	
Environment-related factors	Team Collaboration	Team performance (2)	Team performance (2)
	Tasking	Task type (1)	Task type (3)
		Risk (1)	Risk (8)
		Physical environment (2)	Physical environment (5)
		Multi-tasking requirements (1)	Choice of use (2)
			Clarity of task (1)

TABLE 3.4: Frequency of factors affecting human trust in robots across scenarios in Study 1 and Study 2.

Study 1 (KSA)

Human-related factors . Participants identified **controllability** as a novel ability-based factor in all three scenarios. In scenario 2, two out of 18 participants highlighted the importance of the amount of control humans have in the robots. P1 commented "the reason for trusting the robot in this scenario is that the healthcare practitioner is involved in the process and can intervene to stop the robot if needed". In scenario 1, two participants pointed out the importance of controllability to maintain safe operations. P9 commented "if there is an issue in the robot, the user should take control to stop the robot". One participant in scenario 3 considered familiarity an essential factor. P13 reported "wider use of robots in manufacturing can increase user's trust".

In addition, in line with Hancock et al. [71], two participants mentioned that prior experience with robots will influence human trust in scenarios 1 and 2. In

⁴<https://drive.google.com/drive/folders/1G4v5jDZSxCrEkcQ3MjZKFgNTL-FT4Zkc>

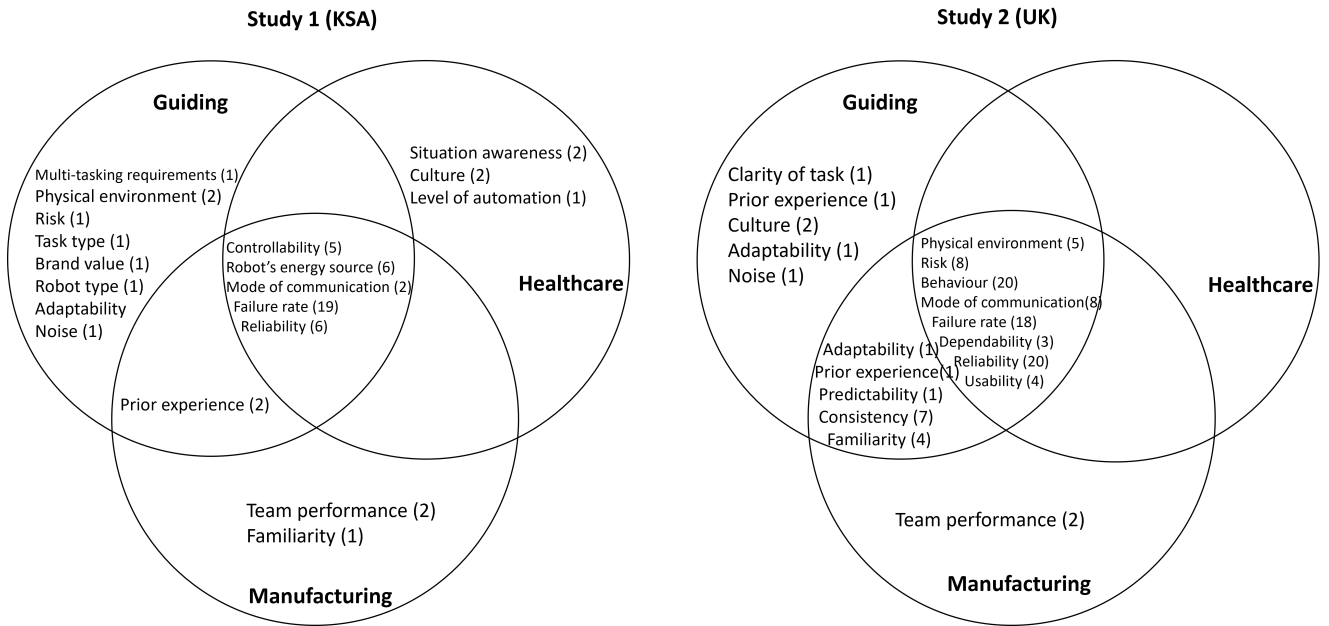


FIGURE 3.4: Commonalities and differences among factors affecting human trust in robots across three scenarios in Studies 1 & 2.

scenario 1, P9 mentioned “the user should have some experience and knowledge about the robot”. In scenario 2, P2 said “the doctor seems to have no experience using the robot, and that could cause trouble”. In Figure 3.4, we can clearly see that factors did vary across different scenarios. For instance, situational awareness was only considered important for scenario 2 by two participants. P14 stated “the community’s awareness could affect the use of robots, especially in healthcare”. P17 mentioned “people understanding robots and how they work in healthcare is important to trust the robot”.

Robot-related factors. Participants identified new performance-based and attribute-based robot-related factors that were not found in the Hancock model of trust [71]. Performance-based factors included noise and the robot’s energy source. Brand value was identified as an attribute-based factor. One participant in scenario 1 considered noise as an important factor. P6 mentioned “noise is one of the factors to be considered because, in this case, the robot is very noisy”. Six participants in scenarios 1 and 2 considered the robot’s energy source as a significant factor affecting user trust. Participants believed that the robot’s battery must be qualified to operate the robot to assist people, particularly

in outdoor environments. For instance, P16 stated “battery life is the critical element while designing such a system: what if the dog ran out of battery, how does the blind person get home?”. P18 also reported the brand value in scenario 1: “the company that makes the robot in the guiding scenario will influence my trust. If Facebook or Google makes the robotic guide robot, my trust will be low because these two companies have a terrible reputation for privacy.” Furthermore, performance-based factors, including mode of communication and failure rate, were commonly presented in all scenarios. Three participants mentioned the mode of communication in all scenarios. For example, P8 stated “in terms of giving feedback, robot dog should communicate verbally to the blind user”. Six participants also presented the behaviour factor in scenarios 1 and 2. Participants described the robots’ behaviour as speed, smoothness of the movements, precision and following of direction. In scenario 1, P2 stated “robot should respond quickly”. P9 mentioned “the robot should choose the correct path”. In scenario 2, P15 reported “the robot’s movement should not be slow” and another mentioned: “the robot’s movement should be fast and precise”. In Figure 3.4, we can clearly see that factors did vary across different scenarios. For example, noise and brand value were only considered in scenario 1. **Environment-related factors.** Risk was the only task-based factor that differed from Hancock’s model of trust [71]. In scenario 1, P6 commented that the robot’s malfunction could lead to serious damage to humans and the environment. In addition, the participants’ responses aligned with the factors identified in [71] regarding team performance, physical environment, and task type. P9 mentioned “the physical environment is essential for trust, and that includes path, object, and the type of connection between humans, and all these elements should be working well”. To compare the three scenarios, Figure 3.4 shows that the factors related to the environment varied across scenarios. For instance, the task type was reported as important in scenario 1. P12 said “the type of task could affect trust even if the robot is the same”. In scenario 3, one participant considered team performance as a team collaboration factor influencing trust. P17 mentioned “the team performance is fantastic, and this will increase the confidence to use and work with robots”.

Study 2 (United Kingdom)

Human-related factors. Participants identified controllability, accountability, the user's condition and familiarity as factors affecting trust in robots. Interestingly, there was no one common human-related factor across the three scenarios. However, in scenarios 2 and 3, five participants considered controllability as extremely important. For instance, P4 said "a human being present with the patient influences the degree of trust that the patient gives to the robot". In scenario 2, one participant felt that the user's condition was important: P5 reported "the patient's health condition can affect the interaction". One participant highlighted accountability as a significant factor: P8 mentioned "in scenario 2, the leader will be responsible for the interaction". In scenarios 1 and 3, four participants highlighted familiarity as an important factor: P11 stated "the everyday use of robotics in a manufacturing setting can lead to a high degree of trust."

Robot-related factors. The findings represent five common factors in all scenarios in line with Hancock's model of trust [71]: behaviour, mode of communication, dependability, failure rate, and reliability. In addition, the analysis identified new performance-based factors, including consistency, usability, noise, and the robot's energy source. Seven participants considered consistency as a critical factor influencing user trust in scenarios 1 and 3. For instance, P11 stated "in scenario 1, I see the consistency is important, does the dog keep moving forwards towards the goal, or does it keep shuffling back?". Four participants reported usability as an important factors. P4 mentioned "has the robot been demonstrated that it's more efficient than human?". Two participants considered the robot's energy source as significant in scenarios 1 and 2. P3 mentioned "the robot's battery should be considered while designing the robot". P6 reported noise in scenario 1 "the dog robot is noisy; that could affect the use and trust". In general, as shown in Figure 3.4, we can clearly see that factors did vary across different scenarios. For example, noise was only considered in scenario 1.

Environment-related factors. The findings identified new ability-based factors, including risk, the clarity of task and the choice of use. The risk factor was common in all scenarios, and eight participants mentioned this. Participants believe autonomous robots can represent more risk to people, mainly when close

to them. For example, P1 reported “in scenario 1, if the robot stops working in a bustling road that is a hazard”. In scenario 1, P9 stated the clarity of the task “the task should be straightforward and align with the object”. P16 also mentioned the choice of use in scenario 2 “choice of use is one of the factors that might influence trust because using robots is not always an option”. Participants’ responses aligned with Hancock’s model [71] regarding team performance, the physical environment and task type. The physical environment was shared in all scenarios. The physical environment includes path, object, and the type of connection between human and robot based on participants’ explanations. The task type was a common factor in scenarios 2 and 3. As we can see in Figure 3.4, environment-related factors varied across the three scenarios. For example, clarity of task was only considered in scenario 1, whereas the choice of use was only in scenario 2.

Results – comparing studies 1 and 2

When comparing the two studies, we find new factors that were not included in [71]. Participants mentioned controllability (59% of the time) followed by familiarity (29% of the time) in both studies as significant human factors affecting their trust in the robot. Interestingly, controllability was considered important in all three scenarios in study 1, but in study 2, it was considered important only in scenarios 2 and 3. Likewise, familiarity was raised once in scenario 3 in study 1 but was cited more frequently in scenarios 1 and 2 in study 2 (see Figure 3.4). For robot-related factors, participants mentioned the robot’s energy source (36% of times), consistency (32% of times) and usability (18% of times). Similar to the human-related factors, participants did not consider these factors in all scenarios. In Study 1, the energy source was highlighted in all scenarios. On the contrary, energy source was cited only in scenarios 1 and 2 in study 2 (see Figure 3.4). Lastly, for environment-related factors, risk (75% of the time) turned out to be the most frequently stated factor affecting a participant’s trust in the robots. Intriguingly, the risk was considered important in scenario 1 in study 1, while in study 2, participants deemed it important in all the scenarios (see Figure 3.4). Table 3.4 further highlights the common factors among the two studies. In summary, the qualitative analysis indicated cultural differences in the two studies when highlighting the importance of factors affecting trust across

different scenarios. In addition, the analysis identified the significance of a factor based on how frequently it was stated by participants; however, we remain conscious that more empirical evidence is needed to establish the significance of a given factor affecting trust. The percentage of the frequency of a new factor was computed by dividing the number of times a new factor was highlighted by the total number of new factors highlighted for a certain type of factor (human, robot, or environment) in both studies.

3.3 Discussion

This study investigated how trust in robots varies across different HRI settings and cultural contexts, yielding several key findings. Our quantitative analysis revealed that while the frequency of trust factors did not differ significantly across scenarios (rejecting H1), trust perception and relevance scores varied significantly by scenario (partially supporting H2) and cultural context (supporting H3). UK participants generally showed higher trust perception scores than KSA participants across all scenarios, while trust relevance varied by scenario type. Our qualitative analysis identified several novel factors affecting trust, including controllability, familiarity, and risk, which varied in importance across different scenarios and cultural contexts. These findings highlight the complex, multi-faceted nature of trust in HRI and the importance of considering both scenario-specific and cultural factors when designing trustworthy robotic systems.

Quantitative findings

H1 indicated that the frequency of factors affecting trust would vary across the three HRI settings. We show that the number of factors (human-, robot-, environment-related) in the two studies did not differ significantly across the scenarios. Hence, H1 was rejected in the light of frequency analysis. Regardless, we see several significant trends and discuss them through the lens of both quantitative and qualitative findings. First, interestingly, robot-related factors were found to be most frequently highlighted in both studies. This finding is comparable with Lewis, Sycara, and Walker [110], which has shown that robot-related factors are the most influencing factors affecting trust in robots. Second,

although the number of factors did not differ across scenarios, the qualitative insights provide evidence that factors affecting users' trust vary across different HRI scenarios. We can clearly see from the analysis that several factors were considered important in one scenario but not in another (see Figure 3.4). For instance, Noise was considered important in scenario 1 in both studies but not in scenario 2 and 3. Lastly, and intriguingly, the number of robot-related and environment-related factors was significantly higher in Study 2 when compared to Study 1. This suggests that participants in the UK stated more and new factors (consistency, usability, and clarity of task) compared to the participants in KSA. All of these findings highlight the multi-facetedness of trust as a construct and reflect on the challenge of measuring it in different HRI settings [1].

H2 predicted that trust in the robots would differ across the three HRI settings. We did see a significant variation of TPS in Study 1 but a non-significant variation of TPS in Study 2. Hence, the results provided partial support for H2a. This finding builds on the existing work that highlights how users' trust ratings vary across different tasks [145, 168, 184]. We see that participants' TPS only differed significantly between scenario 2, scenario 1 and scenario 3 in both studies. It is worth highlighting that scenario 2 dealt with a healthcare context. We speculate that participants were more cautious when trusting robots in healthcare settings. Past work shows that the adoption of robots in healthcare raises performance expectancy, trust, privacy and ethical concerns [122, 10]. It was also intriguing to see that TPS was highest in scenario 3 (manufacturing), followed by scenario 1 (guiding) in both studies. This finding builds on existing literature that has reported how participants assign stereotypes towards robots based on their body shape [28] or their context of use or prior experience [5]. Hence, in this case, participants may have held positive notions about robots in manufacturing or guiding blind users, and this may have resulted in a higher trust score.

The TRS varied significantly across scenarios in both studies. Hence, **H2b** was accepted in both cases. Intriguingly, trust was considered least relevant in scenario 2 in study 1 and scenario 3 in study 2. We speculate that due to the teleoperated nature of the robot in scenario 2, overall, participants may have found it to be less relevant as a human had control of the robot. This was also reflected in their comments, which were described in the qualitative findings. In contrast, we speculate due to participants' backgrounds (living in the UK) and

their prior experience, scenario 3 (manufacturing) was regarded as least relevant in study 2. The adoption of industrial robots in Europe has been significantly higher than in many Arab regions, particularly in sectors such as manufacturing and logistics. According to the International Federation of Robotics (IFR), European countries, including Germany and the UK, rank among the top adopters of industrial robots, with the highest robot density in manufacturing globally [146]. In contrast, while robotics adoption is increasing in Arab regions, many countries are still in the early stages of widespread industrial automation [143, 142]. This suggests that the greater presence of industrial robots in European industries may contribute to higher general exposure and familiarity among individuals in those regions. Comparability, we see a higher level of progression in terms of the use of robots in manufacturing in the UK [54].

The finding further demonstrated compelling trends. In particular, TPS was directly proportional to TRS. It suggests an increase in TPS will cause an increase in TRS or vice versa. Surprisingly, participants in study 1 gave a lower TPS and TRS in scenario 2, and this was in contrast with other scenarios. We understand that participants may have found the healthcare scenario less relevant and, therefore, showed lower trust in the robot. It suggests participants' trust in a robot is dependent on their perception of the relevance (less or more) of the robot in a given setting.

Lastly, **H3** indicated that participants' TPS and TRS would vary significantly across cultures. The analysis confirmed that the TPS and TRS differed significantly across the two different regions. Hence, H3 is accepted. These findings are in line with related work results. (e.g., [178, 20, 112], which also shows that multiple factors associated with an individual's culture affect their perception of trust in robots. Further, the findings show that the TPS scores in all scenarios were higher in Study 2 than in Study 1. Previous studies have shown that participants from Western countries show more trust in robots compared to participants from Eastern countries [73]. In particular, Andrist et al. [20] has shown that Arab participants were more critical of social robots' credibility compared to US participants. These findings also help clarify this trend.

Qualitative findings

We found new factors related to humans, robots and the environment. The human-related factors included controllability, familiarity, the user's condition and accountability. In both studies, participants considered controllability as an important human-related factor in all scenarios except scenario 3 in study 2. We see work in HRI, particularly on measuring or modelling trust in real-time, and found that the amount of control a human operator has in the interaction or the number of times a human takes control during HRI is an indication of their trust [55, 88]. But, interestingly, it is important to understand how much control is sufficient in order to build trustworthy robots and how this varies across different settings [26]. Participants also highlighted accountability as one of the factors. Accountability can be well connected with the amount of control humans will have in autonomous robots. Perhaps this finding helps us think and reflect on the aspect of responsibility in the case of an incident [135, 136].

The robot-related factors included energy source, consistency, usability, noise and brand value. The environment-related factors included task clarity, choice of use, and risk. Brand value was the only new factor in Study 1. We found that individuals belonging to Arab culture care more about brands than Europeans [13] and believe that this explains the given finding. The risk factor was common in both studies. Participants believed that the degree of risk involved in a task can affect their trust. For example, in the guide robot scenario, all participants in both studies reported risk because the blind human was in a vulnerable situation. Other factors identified the vulnerability of such a robot's energy source. For instance, participants were worried about the battery life of the robot in the guide robot scenario. Since we analyse our data in the light of Hancock et al.'s model [71], we consider the risk to be a new factor. It is interesting to note that recent work also reflects on the role of risk in human-robot trust [171, 72]. We also see common factors in studies 1 and 2 that can be seen in Hancock et al.'s model of trust [71] (see Table 3.4). Reliability and failure rates were presented in both scenarios. According to Washburn et al. [180], reliability and error rate factors are related to each other and have a strong relationship with human trust. All participants in both cultures stated that the mode of communication in all HRIs was a significant factor. For example, participants suggested that the robot communicate verbally since the user is blind. We posit

that when the user has a proper way to communicate with the robot and receives clear feedback, the level of trust will increase significantly.

3.4 Conclusions, limitations and future work

This chapter has explored the factors affecting human trust in robots across different HRI settings and cultural contexts. Our findings demonstrate that trust is a complex, multi-faceted construct that varies significantly based on both the specific interaction scenario and cultural background. We identified several novel factors affecting trust, including controllability, familiarity, and risk, which were not fully explored in previous literature.

The insights gained from this study directly inform the development of our computational trust model in Chapter 4. By understanding how factors such as controllability, risk, and cultural background influence trust perception, we can create a more robust and adaptable model that accounts for the dynamic nature of trust across various HRI contexts. The identification of these factors ensures that our model incorporates the most relevant variables for predicting and calibrating trust in different scenarios.

Furthermore, our cross-cultural comparison between KSA and UK participants highlights the importance of considering cultural differences when designing trustworthy robotic systems. The significant variations in trust perception and relevance scores between the two cultural groups emphasize the need for culturally sensitive approaches to trust calibration in HRI.

Despite the valuable insights gained from this study, there are certain limitations to acknowledge. First, the order of scenario presentation was not counterbalanced, meaning all participants viewed the scenarios in the same sequence. This could have introduced order effects, where responses to later scenarios may have been influenced by prior exposure to earlier ones. Future studies should consider randomising scenario order to minimise potential biases. Second, all participants were recruited from computer science departments. While this ensured familiarity with robotics, which was essential for identifying trust-related factors, it also presents a limitation regarding generalisability. The findings may not fully represent the perspectives of individuals from other

disciplines or those with limited exposure to robotics. Future research should include participants from a more diverse range of backgrounds to gain broader insights into trust dynamics in HRI.

By providing a comprehensive analysis of the factors affecting trust in diverse HRI settings, this chapter establishes the foundation for the mathematical trust model developed in subsequent chapters, ensuring that it accurately reflects the complex dynamics of human-robot trust in real-world applications.

Chapter 4

Mathematical Trust Model

This chapter¹ attempts to bridge the gap in the limited work on creating mathematical models that capture the complexities of human trust in robots across diverse interaction contexts. To create the model, we utilised the findings of chapter 3, which emphasise the importance of trust-related factors—such as controllability, risk, and robot reliability and how these factors can vary across different interactions. The model aims to provide a structured framework for capturing and dynamically adjusting trust during repeated interactions. Trust in HRI is a complex construct that influences user reliance and the quality of interactions. Previous models have often been limited to specific settings or mainly addressed collaborative contexts [92, 164]. In collaborative contexts, robot trustworthiness is crucial to enable humans and robots to work together safely and efficiently in shared workspaces [116]. However, competitive interactions, where important elements like robot truthfulness come into play, remain underexplored despite their increasing relevance in scenarios where humans and robots may compete against each other [106]. We have attempted

¹This Chapter has been published as two conference papers:

- Ahmad, M., Alzahrani, A., Robinson, S., & Rahat, A. (2023, December). Modelling Human Trust in Robots During Repeated Interactions. In Proceedings of the 11th International Conference on Human-Agent Interaction (pp. 281-290).
- Alzahrani, A., & Ahmad, M. (2024, March). An Estimation of Three-Layered Human's Trust in Robots. In Companion Proceedings of the 29th International Conference on Intelligent User Interfaces (pp. 144-146).

Also, a part of this chapter has been submitted to a journal, and it is under review:

- Abdullah Alzahrani, Simon Robinson and Muneeb Ahmad. A Three-Layered Framework for Estimating Human Trust in Robots During Repeated Interactions. International Journal of Social Robotics. [Under-Review].

to bridge this gap by developing a more generalisable mathematical model for both competitive and collaborative experimental settings.

To build the model, we draw on insights from the theoretical model proposed by Hoff and Bashir [78], which distinguishes between dispositional, situational, and dynamically learned trust. To ensure the model's adaptability across different contexts, we validate it through studies conducted in both competitive and collaborative HRI settings.

Initially, we developed a baseline trust model and validated it through the first study in a competitive setting. Subsequently, based on insights gained from this study, we refined the model to better capture the dynamics of trust in collaborative settings and validated it through two additional studies. By using this iterative approach, we have established a robust, adaptable trust model suited to varied HRI contexts.

The objectives of this chapter are as follows:

- Develop a mathematical trust model incorporating dispositional, situational, and dynamically learned trust.
- Validate the model through empirical studies across competitive and collaborative HRI settings to ensure it can accurately capture real-time trust variations.
- Explore how these trust layers interact and influence each other over repeated interactions.

In addressing these objectives, this chapter explores the following research questions:

- **RQ1:** How can we effectively model and validate the three layers of trust (dispositional, situational, and dynamically learned) across competitive and collaborative HRI settings, building on the findings from Chapter 3?
- **RQ2:** How does dynamically learned trust evolve over time during repeated HRI in different contexts (competitive and collaborative, building on the findings from Chapter 3)?
- **RQ3:** How do dispositional, situational, and dynamically learned trust layers influence each other over repeated HRI?

These research questions directly address the overarching aim of this thesis: to develop a real-time computational model for measuring human trust in robots during interaction. RQ1 focuses on effectively modeling and validating the three-layered trust framework, which forms the foundation of our computational approach to measuring trust in real-time. By distinguishing between dispositional, situational, and dynamically learned trust, we can capture both the static and dynamic aspects of trust formation during human-robot interaction. RQ2 examines how trust evolves over time during repeated interactions, which is essential for developing a model that can adapt to changing trust levels in real-time. RQ3 investigates the interrelationships between the three trust layers, which is crucial for understanding how initial trust dispositions and situational factors influence the dynamic learning of trust over time. This understanding enables our computational model to account for individual differences and contextual variations when measuring trust in real-time.

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

1. **Contribution 1:** We present a novel mathematical model for measuring human trust toward robots in real-time during interaction, integrating dispositional, situational, and dynamically learned trust layers. This model advances beyond existing approaches by computing trust during the interaction itself rather than retrospectively, and we validate it across three different repeated HRI settings.
2. **Contribution 2:** Using the Trust Perception Scale (TPS) questionnaire [159], we show that the model predictions of trust scores strongly agree with the trust perception of participants in an experiment where they interacted with an NAO robot in a *novel trust-based game*.
3. **Contribution 3:** We show that our dynamically learned trust model, which adapts to changing interaction contexts and user behaviours, captures significant trust variations over time, with the most notable differences occurring during the final interaction with the robot. This demonstrates the model's ability to detect meaningful shifts in trust levels as relationships evolve.

4. **Contribution 4:** We show that there exists a strong positive correlation between different layers of dynamically learned trust across different HRI settings.

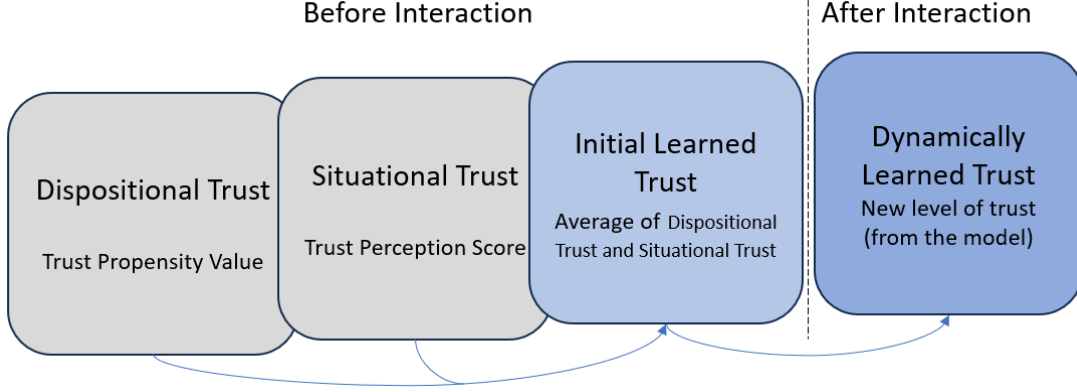


FIGURE 4.1: Modelling the Three Layers of Trust.

4.1 Initial Trust Model

We begin our development of a computational trust model with a competitive game scenario. This approach allows us to establish baseline trust dynamics in a controlled environment before extending the model to more complex interaction scenarios. The competitive game provides a structured context where trust decisions have clear consequences, making it an ideal starting point for our initial trust model.

We focus first on the **dynamically learned trust** as depicted in figure 4.1, which 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 [78].

With these interpretations, inspired by the experiential model proposed in Jonker and Treur [85], we relate different layers of trust as follows:

$$T(t + \Delta t) = T(t) + \gamma(E(t) - T(t))\Delta t, \quad (4.1)$$

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 t th interaction, and $T(0)$ is the initial

trust at $t = 0$, i.e. when no interactions have occurred. Here, Δt represents the unit difference between events. Thus, $\Delta t = 1$. The parameter γ is the learning rate, which determines the weight assigned to new experiences in updating trust over time.

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.

This means that the magnitude of the difference between the current trust level and the experience from the latest interaction determines whether the new trust level should increase, remain the same, or decrease.

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

$$E(t) = \begin{cases} \sum_{i=1}^t \frac{P_i C_i}{K} & \text{if } K > 0. \\ 1 & \text{if } K = 0. \end{cases} \quad (4.2)$$

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 i th 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, refers to the total instances in which the participant disagrees with or challenges the robot's decisions during

the interaction. 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 chapter in Section 4.2.2.

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 was clear that $T(t) \in [0, 1]$ with 1 representing a complete trust, and 0 illustrating a complete distrust; see Figure 4.2. It was, therefore, reasonable to consider an initial trust of $T(0) = 0.5$ which means that the human has neutral trust at time point 0 [79], in the absence of some high-level ancillary information. We use this value of $T(0)$ throughout this chapter.

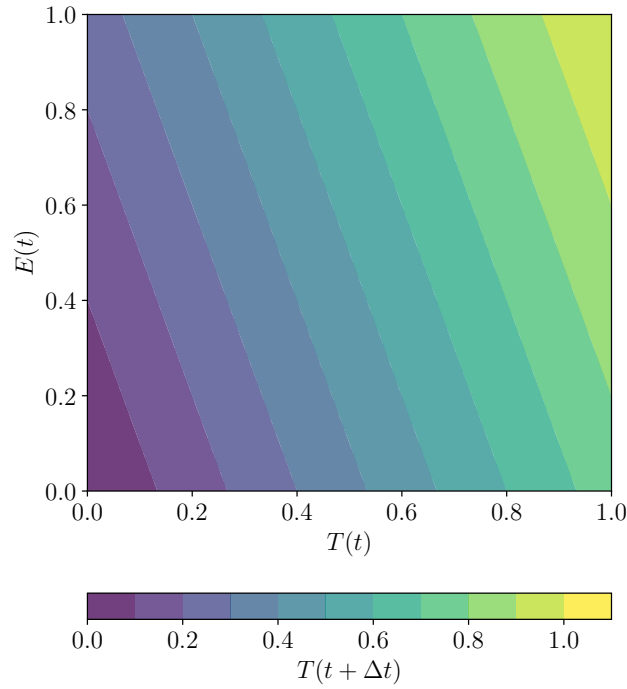


FIGURE 4.2: An illustration of how the values of $T(t)$ and $E(t)$ impact $T(t + \Delta t)$ given $\lambda = 0.25$. Unsurprisingly, when trust is low, an immediate, highly positive experience does not alter learned trust substantially.

4.2 Formative Evaluation of Approach

The first study in this chapter was designed to validate the initial mathematical trust model and involved participants in interacting 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 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.

4.2.1 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]. Participants were provided with a participant information sheet outlining the study objectives, procedures, and data confidentiality measures. Prior to participation, they gave informed consent by signing a digital consent form, which explicitly stated their right to withdraw at any time without consequences and without providing any reason. Participants were informed that they could withdraw their data up to two weeks after participation by contacting the researchers. The study adhered to institutional ethical guidelines for research involving human participants, including principles of voluntary participation, informed consent, and data protection.

4.2.2 System description

The system shown in Figure 4.3 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

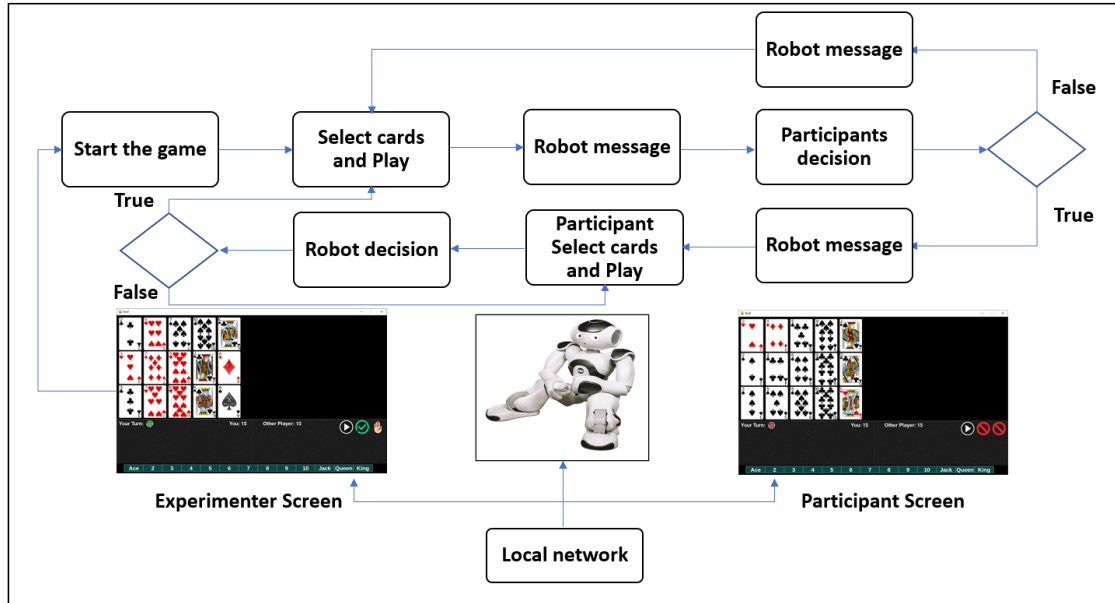


FIGURE 4.3: System architecture for the trust measurement platform. The system integrates multiple components: (1) a card game module that creates trust-based decision scenarios, (2) a semi-autonomous NAO robot that interacts with participants, (3) a trust computation module implementing our mathematical model, and (4) a data collection system capturing interaction patterns. The architecture enables real-time trust measurement during HRI in a competitive game setting.

participants. The goal of 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.

The Game

We developed an interactive two-player card game, *Bluff Game*, using Python, that participants can play against the robot as shown in figure 4.4. 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). The decision to follow or reject the robot’s suggestions serves as a behavioural trust measure, which is well-established in the literature. According to Chita-Tegmark et al. [36], trust in HRI can be measured through behavioural tasks, where users’ actions directly indicate their trust levels. Similarly, Gaudiello et al. [61] demonstrated that trust serves as a direct indicator of robot functional acceptance, where users

choose whether to conform to robot suggestions. 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 the 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 less than the opponent's 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 4.1). 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 $P_i \leftarrow \neg B_i$, i.e. $P_i = 0$ if the robot was truly bluffing and *vice-versa*. We use this in (4.1) to update trust in the first experiments in this chapter.

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

TABLE 4.1: Truth table of B_i , C_i and P_i at the i th interaction.

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. This strategy is fully documented in the Appendix A, which details the exact sequence of robot suggestions, including the timing, accuracy, and complexity of each suggestion. 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.

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 occasions, the robot thanked the participants 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 starts, the NAO robot informs the participant that "the game starts now". The 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, "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.

4.2.3 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*²).

Participants were classified as experienced with robots as 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.

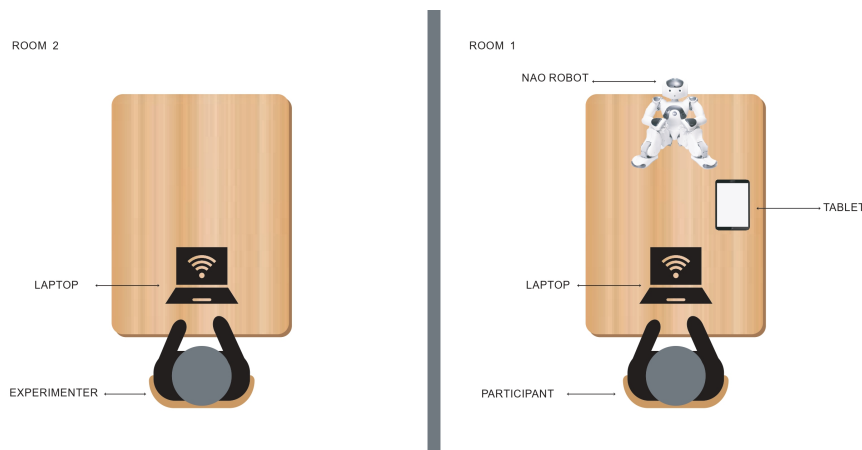


FIGURE 4.4: Experiment Setup - it depicts an experimenter controlling the robot in a one room (left), while participant playing the game against the robot in another room (right).

4.2.4 Setup and Materials

We conducted this study in 2 separate rooms, as shown in Figure 4.4. 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. 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.

²<https://calendly.com>

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

4.2.5 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.
3. Participants wore glasses and a wristband to monitor physiological responses during the interaction. It should be noted that the data collected from these devices has not been analysed in the current study. This physiological data will be analysed in Chapter 5.
4. The 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 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 would receive a £10 Amazon voucher for their participation in the study.

4.2.6 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 [159]. 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-item subscale because it helps measure changes in trust over time and during multiple trials. Following [159], 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 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.

Statistical analysis was conducted using SPSS. Specifically, we employed the multiple linear regression and repeated measure ANOVA.

4.3 Results of Trust Model Formative Evaluation

This section presents the results of our formative evaluation of the initial trust model in a competitive game setting. We analyzed both quantitative and qualitative data to assess the model's accuracy and alignment with human trust perception. The quantitative analysis focused on three main aspects: (1) the relationship between Trust Perception Scale (TPS) scores and Trust Modelled Scores (TMS), (2) the effect of repeated interaction sessions on trust development, and (3) the dynamics of trust change over time. We also examined how risk, control, and robot performance influenced trust across different interaction sessions. Below, we present detailed findings for each of our three hypotheses.

Question format: “To what extent do you agree that the robot was...”
 Participants responded on a continuous scale from 0% (“Not at all”) to 100% (“Completely”).

Item No.	Statement
1	Was dependable.
2	Was reliable.
3	Was predictable.
4	Acted consistently.
5	Functioned successfully.
6	Met the needs of the mission/task.
7	Provided appropriate information.
8	Communicated with people.
9	Provided feedback.
10	Followed directions.
11	Performed exactly as instructed.
12	Had errors. [Reverse coded]
13	Was unresponsive. [Reverse coded]
14	Malfunctioned. [Reverse coded]

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

TABLE 4.2: Trust Perception Scale (TPS) 14-item Subscale

To test **H1**, a multiple linear regression was calculated to predict TMS based on TPS and session. Multiple linear regression is appropriate for examining the relationship between a dependent variable and multiple predictors, allowing us to assess the contribution of each factor while controlling for others [128]. A significant regression equation was found ($F(2,177) = 7.36, p < .001$), with an R^2 of .077 (see figure 4.5). 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 session. We understand that trust in a robotic system changes based on the perceived performance (truthfulness or error rate) [164, 78, 71]. 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 [138, 99]. Perceived risk refers to an individual’s feeling that a specific task or context has

potential negative outcomes [170]. In this case, when the participants' number of cards left was significantly higher than the robots' number of cards left, it presented a high risk of losing the game. Hence, it 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 4.2.6) 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 [55, 167, 71, 91]. 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 show that factors affecting trust may impact differently across different situations, as shown in the findings of [18]. Lastly, we did not find a relationship between TPS, risk, performance or control in all four sessions.

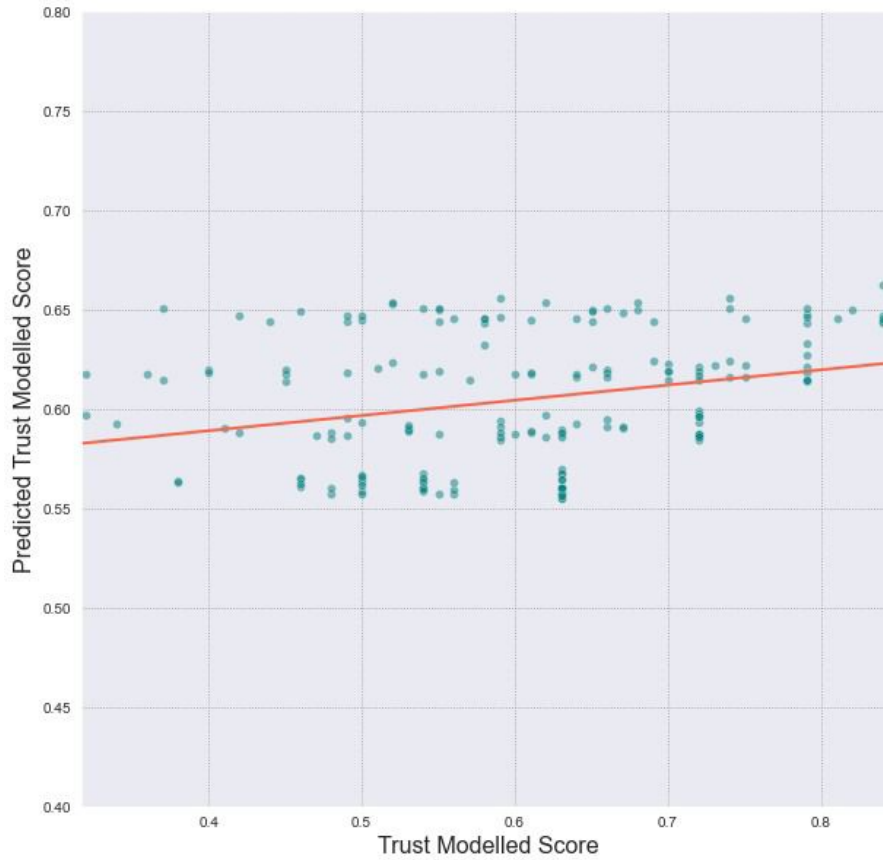


FIGURE 4.5: Scatter plot and linear regression line showing a relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and time.

To test **H2** and **H3**, a repeated-measures ANOVA, which is suitable for analysing changes in trust scores over time while accounting for within-subject variability [52] was conducted to determine whether there is an effect of the interactive session (session 1, session 2, session 3, and session 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.

Post hoc Bonferroni correction was applied to adjust for multiple comparisons, reducing the risk of Type I errors and ensuring the reliability of significant differences [130]. The comparison 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 sessions 3 and 4, respectively. The mean and Standard deviation for both TPS and TMS can be seen in Table 4.3.

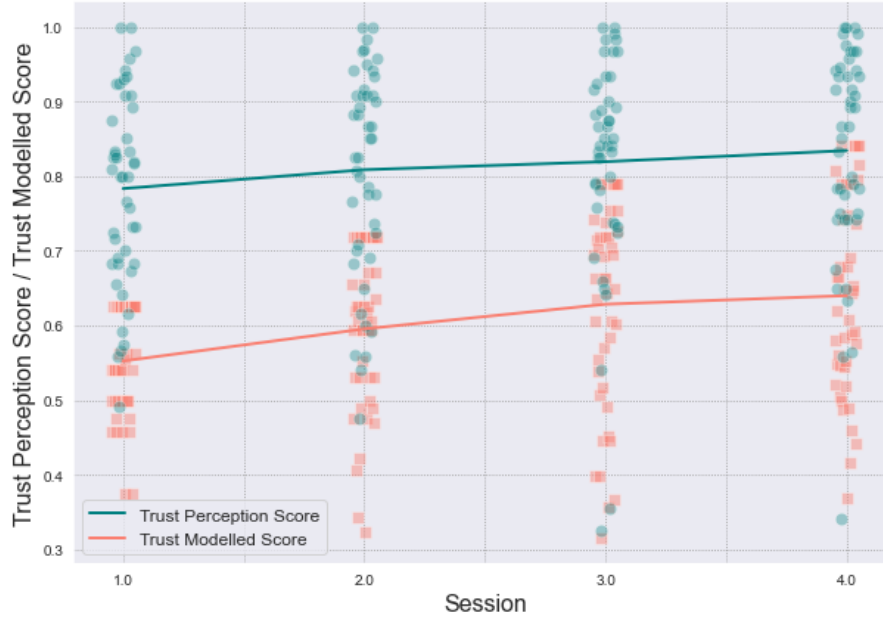


FIGURE 4.6: Scatter plot illustrating the evolution of trust over four interactive sessions. Trust Perception Score (TPS) is shown in green, representing participants' subjective trust ratings, while Trust Modelled Score (TMS) is shown in orange, representing the trust computed using the mathematical model. The trend lines indicate a gradual increase in both TPS and TMS, suggesting that trust develops positively over time.

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 the 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 [43], 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, especially in a competitive setting where we used truthfulness as an indicator of trust,

is novel. In competitive interactions, including the Bluff Game, trust is not only about expecting absolute honesty but also about predictability and rational decision-making. In any competition, trust is necessary for decision-making, as players do not immediately assume deception. Instead, they assess the opponent's behaviour over time. For instance, in adversarial negotiation or strategic military simulations, a trustworthy opponent is one whose behaviour can be analysed and responded to, rather than one that operates in a totally inconsistent or deceptive manner.

The presented work is the first effort, and we 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 future changes to the model based on this work.

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 sessions ($r = 0.65, p < .01$ for T_1 , $r = 0.72, p < .001$ for T_2 , $r = 0.68, p < .01$ for T_3 , and $r = 0.75, p < .001$ for T_4). These results suggest a positive increase in dynamically learned trust across the four different sessions, confirming the alignment between subjective trust perception and the modelled trust score.

4.3.1 Discussion and Implications of Formative Evaluation

The formative evaluation of our initial trust model yielded several important insights. First, while the model successfully captured general trust trends, it struggled to account for individual differences in trust calibration rates. Second, the model's performance varied across different interaction phases.

These findings highlight the need for an extended model that can incorporate other factors, such as risk and ambiguity, and be tested in a different context. The extended model developed in response to these limitations incorporates these

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 4.3: Mean (M) and standard deviation (SD) of TMS and TPS scores across the four sessions. The Shapiro-Wilk test confirmed that both TMS ($W = 0.967, p = 0.085$) and TPS ($W = 0.971, p = 0.310$) follow a normal distribution, justifying the use of mean as a central measure.

trust parameters and contextual weighting factors, addressing the core thesis aim of measuring trust in real-time across diverse interaction contexts.

The positive correlation between TPS and TMS across all four sessions validates our approach of using a computational model to estimate human trust in robots. However, the lack of significant correlation between TPS and factors like risk, performance, and control suggests that subjective trust perception involves complex psychological processes not fully captured by our initial model. This insight directly informs the development of our extended model, which incorporates additional factors and a more sophisticated mathematical framework.

This progression from initial to extended model demonstrates the iterative refinement process central to developing robust trust measurement tools for HRI, advancing the thesis goal of creating a comprehensive framework for real-time trust measurement.

4.4 Extended model

The extended version of the model builds on insights from the initial study. The finding from the first study demonstrated that perceived risk significantly influenced participants' trust in robots across repeated HRI contexts. Additionally, decisions in HRI often involve a degree of uncertainty, as it is challenging to fully predict the outcomes of a robot's actions. By refining the model to incorporate risk and uncertainty, we aim to create a framework

that more accurately reflects how trust evolves in complex, real-world scenarios. Furthermore, to achieve a comprehensive measure of trust, this model includes three distinct layers: dispositional, situational, and learned trust, as shown in Figure 4.1. In this approach, we have chosen specific scales to compute these layers, aligning with the best practices in trust measurement within HRI as detailed by Krausman et al. [100].

For computing **dispositional trust (DT)** values, we utilised a Likert scale questionnaire [47]. We computed the **situational trust (ST)** value using the trust perception scale [158]. The initial trust can be better reflected by averaging propensity and situational trust, which considers past pre-interaction experiences with the system. Therefore, we considered the **initial learned trust** $T(0)$ as the average of dispositional and situational trust:

$$T(0) = \frac{DT + ST}{2}. \quad (4.3)$$

The rationale for this approach is that both dispositional and situational trust, as pre-interaction stages, contribute equally to shaping the user's initial expectations and trust levels before any direct interaction with the robot. Dispositional trust offers a stable baseline, reflecting an individual's inherent tendency to trust, while situational trust modifies this baseline based on the specific context and conditions of the interaction. By averaging these two components, the initial trust calculation captures both the enduring personal characteristics and the dynamic environmental factors, providing a more balanced measure of the user's initial trust.

To compute **dynamically learned trust**, we used the same equation 4.1 in the initial model but with differences in the experience computation as:

In this refined version of the model, the experience $E(t)$ is calculated based on human decision-making behaviour, the performance of robots, risk, and ambiguity aversion in a given task as follows:

$$E(t) = (1 - (\frac{\sum_{i=1}^N |P_i C_i - C_i R_i|}{N})) - A(t) \quad (4.4)$$

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

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

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

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

$$A(t) = \frac{\sum_{i=1}^N |K_i - F_i|}{N}, \quad (4.5)$$

Where K_i is the expected number of robot failures (how many times the user overrides the robot), F_i is the actual number of robot failures at time t , and N is the total number of instances. With this representing of $E(t) \in [0, 1] \subset \mathbb{N}$, and an initial $T(0) \in [0, 1] \subset \mathbb{N}$, it is clear that $T(t) \in [0, 1]$ with 1 representing a complete trust, and 0 illustrating a complete distrust; see Figure 4.7.

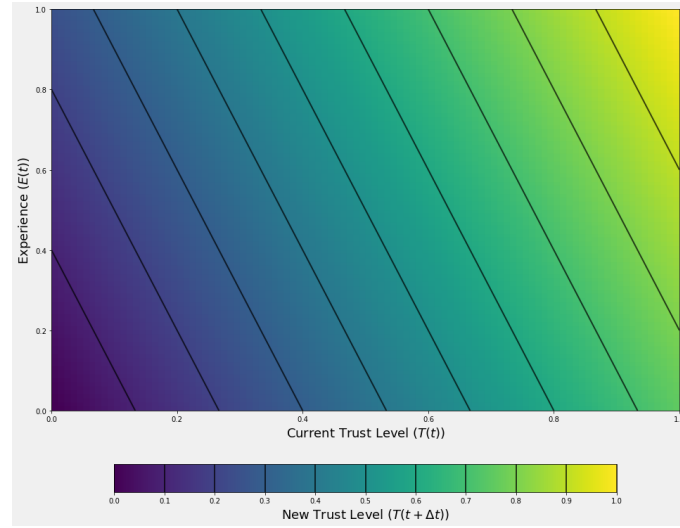


FIGURE 4.7: Illustration of the impact of Current Trust Levels $T(t)$ and Experiences $E(t)$ on the New Trust Level $T(t + \Delta t)$ for $\gamma = 0.25$, showing that a highly positive experience has a limited impact when current trust is low.

4.5 Summative Evaluation of the Model

We designed a study to validate the enhanced mathematical trust model, involving participants interacting with the NAO robot on four different occasions during collaborative HRI, with each session lasting approximately 7.45 minutes. Each session contained multiple decision points where participants had to decide whether to accept or reject the robot's suggestions. At each decision point, the model computed instantaneous trust, dynamically updating it throughout the session based on these interactions. By the end of each session, the cumulative experience, combined with the previous trust score, formed a new trust level. After each session, participants completed a questionnaire to assess their perceived trust in the robot. This setup allowed us to compare the model's real-time computed trust scores with participants' self-reported trust levels. All participants followed the same sequence of four interactive sessions to ensure consistency in the study conditions. While randomisation is often used in such studies, we chose a fixed session order to focus on measuring trust dynamics over time. This uniform approach allowed us to observe trust evolution consistently across participants. All sessions occurred on the same day, with a 5-minute interval between sessions.

We tested the following hypotheses:

- **H4:** Both the TPS and interaction session (time) will predict the TMS in a collaborative HRI setting.
- **H5:** We will observe a significant interaction effect on sessions (session 1, session 2, session 3, and session 4) for TMS and TPS scores, reflecting that humans dynamically learned trust in robots will change over time during repeated HRI in a collaborative setting.
- **H6:** We will observe variations in the interplay or correlation among the three layers of trust – dispositional, situational, and learned (both initial and dynamic).

These hypotheses directly address the central aim of this thesis: to develop a real-time computational model for measuring human trust in robots during interaction. H4 tests whether our enhanced model can accurately predict trust in collaborative settings, extending the validation from competitive contexts and supporting our goal of creating a generalizable trust measurement approach. H5 examines the dynamic nature of trust over repeated interactions, which is essential for developing a model that can adapt to changing trust levels in real-time—a core requirement for effective trust measurement during ongoing human-robot interaction. H6 investigates the relationships between different trust layers, providing insights into how initial dispositions and situational factors influence trust development over time, which is crucial for creating a comprehensive trust measurement framework that accounts for individual differences and contextual variations.

4.5.1 Ethics

The study was submitted for ethical review and was approved by the university ethics board. Approval number: 2202370516013. Participants were provided with a participant information sheet outlining the study objectives, procedures, and data confidentiality measures. Prior to participation, they gave informed consent by signing a digital consent form, which explicitly stated their right to withdraw at any time without consequences and without providing any reason. Participants were informed that they could withdraw their data up to two weeks after participation by contacting the researchers. The study adhered

to institutional ethical guidelines for research involving human participants, including principles of voluntary participation, informed consent, and data protection.

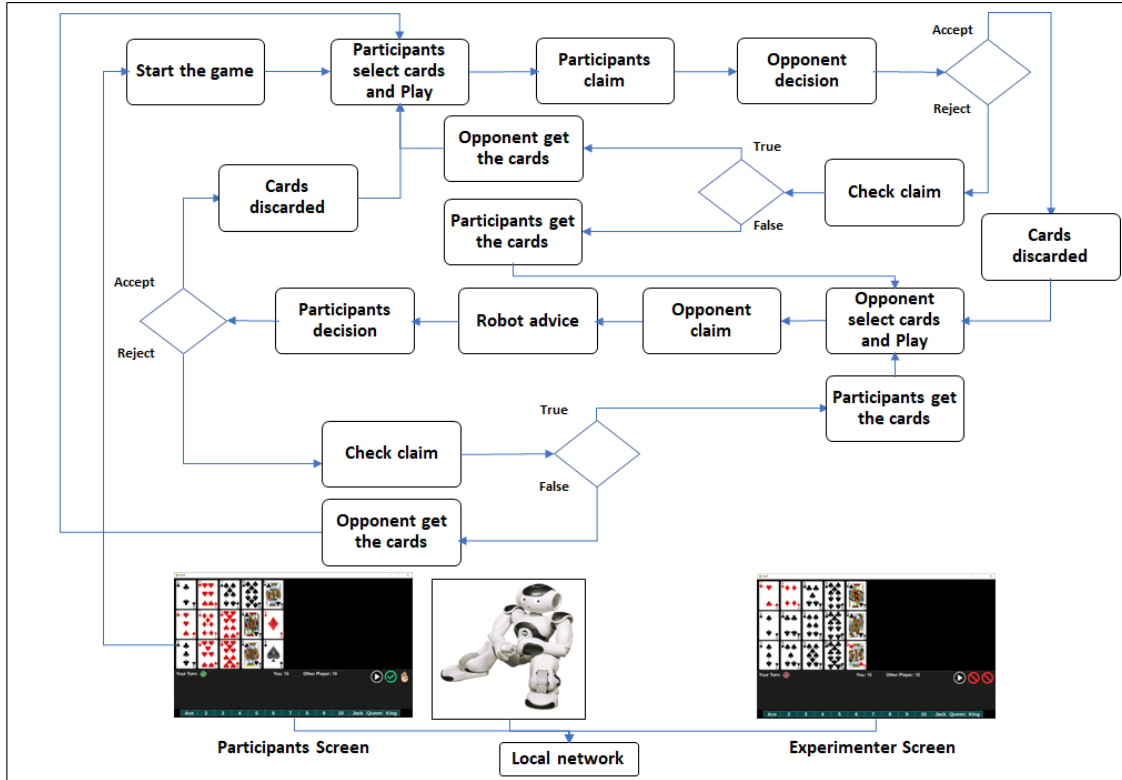


FIGURE 4.8: Overview of the trust-adaptive system. The system comprises two integrated modules: (1) an interactive Bluff Card Game designed to elicit varying levels of trust and distrust by placing participants in decision-making scenarios under conditions of uncertainty and risk, and (2) a semi-autonomous robotic partner that provides advice and engages dynamically with the participant during gameplay. The system captures behavioural, physiological, and decision-based responses to investigate the influence of robot advice accuracy, participant control, and perceived risk on trust formation and calibration. This setup enables real-time assessment and modelling of human trust dynamics in repeated HRI.

4.5.2 System description

The system presented in Figure 4.8 consists of two modules. The first module is an interactive card game that generates various situations for participants to either trust or distrust the robot. The second module is a semi-autonomous

robot that plays the game with the participants and assists them in making decisions. The model is designed to estimate human trust in the trustworthiness of the robot, particularly in situations that present risk and uncertainty. In the Bluff Game, we focus on key factors that impact trust, such as the robot's accuracy in providing advice, the participant's control in accepting or rejecting the robot's advice, and perceived risk (when the player's cards are more than the opponent's), which is indicated by the proportion of the participant's cards to the opponent's cards. The main objective of the system is to analyze the participants' reactions to situations that involve trust with the robot and how the robot's behaviour over time impacts their decisions to trust it.

The Game

In this study, we employed the **Bluff Game** that we developed earlier. We modified the game to create trust and distrust scenarios in a collaborative setting. 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](#) and in the Appendix [A](#). 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 [[188](#), [62](#), [5](#)].

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

The game's dynamics are specifically designed to incorporate factors such as risk and ambiguity, which are integral to the conceptual framework of trust. Risk in the game arises when a player has significantly more cards than their opponent. Additionally, the game involves an element of uncertainty due to the ambiguity of the robot's advice, challenging players to navigate decisions under ambiguous conditions. This aspect is crucial for reflecting the complexity

and unpredictability present in HRC, effectively simulating real-world scenarios where decisions must be made with incomplete information.

The calculation of experience $E(t)$ and the dynamically learned trust in the game setting hinges on several key variables. Risk, which can be defined in HRI as an individual's perception of the possible negative consequences associated with interacting with robots [139]. This perception is based on their knowledge and experience of the task, regardless of their personal history or familiarity with the system, technology or person that may be involved in that situation [139]. In this context, Risk was quantified as the risk index R_i . Specifically, R_i is given a value of 1 if Player 2 has more cards than Player 1, which directly impacts the perceived likelihood of negative outcomes (losing the game) if unable to eliminate their cards first. Otherwise, R_i is assigned a value of 0.

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

The term $|P_i C_i - C_i R_i|$ will represent the player's behaviour by aligning the robot's performance and the participants' control, and incorporating the associated risks during the game (see 4.5). The truth table indicates a value of 1, showing misalignment, in two scenarios: when performance is low, but control and risk are high $P_i = 0, C_i = 1, R_i = 1$, and when performance is high, control is high, but the risk is low $P_i = 1, C_i = 1, R_i = 0$. A value of 0, indicating alignment or no control by the user regardless of the risk level, applies in all other situations. This differentiation is crucial for accurately calculating the experience $E(t)$ within various risk contexts.

Ambiguity, in this context, refers to situations where the outcome of following the robot's advice was not immediately clear or predictable. For example, the robot might suggest accepting the opponent's claim, but if that claim turned out to be false, the cards would be discarded without revealing their true value. Ambiguity aversion was applied in the following manner: $A(t)$ reflects the user's aversion to uncertainty surrounding the robot's performance. A difference between K_i and F_i in each instance indicates a mismatch between the expected and actual robot performance, contributing to the overall Ambiguity Aversion $A(t)$. This metric is important to understand the influence of the user's uncertainty on their instantaneous trust (experience) in the robot during the game. (see Table 4.5).

P_i	C_i	R_i	$ P_i C_i - C_i R_i $
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

TABLE 4.4: Truth Table for $|P_i C_i - C_i R_i|$

K_i	F_i	$ K_i - F_i $
0	0	0
1	0	1
0	1	1
1	1	0

TABLE 4.5: Truth Table for $|K_i - F_i|$

Interaction Scenarios

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

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

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

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

6. If the participants asked the robot to repeat the suggestion, the robot repeated the suggestion for them.
7. If the robot did not hear the participants, the robot said "Sorry, I did not hear that, could you please repeat it".
8. If the participants seem to have been occupied with something else, the robot said "You seem occupied with something else, could you please focus on the game".
9. If the participants asked the robot for anything else during the game, the robot said "I can only advise you when you are deciding to accept or reject".

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

4.5.3 Participants

This study was conducted with 45 participants aged between 18 and 40 years. The age distribution averaged 33.13 years with a standard deviation of 6.22. Out of the 45 participants, 19 were females, 25 were males, and one participant chose not to disclose their gender. We invited participants to partake in the study via the university's electronic mailing system and flyers around the university campus. Participants were able to book their slots for the study using the online scheduling platform *Calendly*³.

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

³<https://calendly.com>

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

4.5.4 Setup and Materials

In the study, we utilised two separate rooms, as illustrated in Figure 4.9. In the first room, the participants had a laptop to play the game while the robot was positioned on the table next to them. The participants were seated beside the robot. The participants used a tablet to complete questionnaires before and after each game round. In the second room, the experimenter sat in front of a laptop to control the game, robot, and overall interaction.

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

4.5.5 Procedure

The study was conducted in the following steps:

1. On entering the lab, participants were greeted by the researcher and completed the propensity to trust questionnaire before proceeding with the study.
2. Participants received the experiment information sheet and game instruction sheet and signed the consent form.
3. Participants completed the demographics questionnaire, including information about their experience with the robot.



FIGURE 4.9: Experiment Setup. An experimenter controls the robot in one room (left), while the participant is playing the game with the assistance of the robot in another room (right).

4. Participants were given a demonstration of the game and had time to practice, allowing for a better understanding of the game and the interaction with the robot.
5. Participants completed the pre-interaction questionnaire.
6. Participants wore glasses and a wristband, and the experimenter began recording the data to be collected from these devices and left the room.
7. Participants engaged in the game alongside the NAO robot, with their interactions being recorded, while the experimenter remotely controlled the gameplay and robot from the other room.
8. After each game, the experimenter walked into the room, asked the participants to complete the post-interaction questionnaire.
9. The rest of the study repeated steps 6, 7, and 8 on three different occasions.
10. At the end, participants were thanked for their participation and were told that they would receive a £10 Amazon voucher as a token of appreciation for their participation in the study.

4.5.6 Measurements

In this study, we expand our measurements to assess pre-interaction human trust and calculate both dispositional and situational trust.

- Before participating in any interaction or gaining awareness of the surrounding environment of the interaction, the participants were asked to complete a 10-item questionnaire on the tablet to assess dispositional trust [47]. This questionnaire utilised a 5-point Likert scale ranging from "Strongly Agree" to "Strongly Disagree" for responses. The items on the questionnaire are detailed in Table 4.6. This scale is a recently developed tool that benefits from up-to-date insights into trust, making it highly relevant for modern contexts like HRI. It was developed through the rigorous Delphi method, involving multiple rounds of expert feedback, ensuring strong content validity and a balanced focus on both trust and distrust, which is essential for capturing different trust behaviours in HRI. Additionally, its transparency and interdisciplinary input make it more robust than older scales, which may lack such comprehensive development.
- After becoming aware of the interaction and the role of the robot, but before the primary interaction, participants completed a pre-interaction questionnaire to assess their situational trust towards the robot by rating the robot on the TPS scale [159].
- To validate the model's credibility, we used TPS subjective measures of trust, similar to our approach in the first study.

4.6 Results of Summative Evaluation

This section presents the results of our summative evaluation of the enhanced trust model in a collaborative setting. We analysed the performance of the model on three hypotheses: (1) the predictive relationship between the trust perception scale (TPS), the interaction session, and the trust modeled score (TMS), (2) the effect of repeated interactions on trust development, and (3) the relationships between different trust layers. We also compared the performance of our initial and enhanced models to validate the improvements made. The analysis includes

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

TABLE 4.6: Dispositional Trust Questionnaire Items

multiple linear regression, repeated-measures ANOVA, correlation analysis, and comparative model evaluation. Below, we present detailed findings for each hypothesis.

4.6.1 H1: Predicting TMS with TPS and Session

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

The regression model was found to be highly significant, $F(2,177) = 265.605, p < .001$, with $R^2 = 0.750$ (Adjusted $R^2 = 0.747$), meaning that 75% of the variance in TMS is explained by TPS and Session Variables (see Figure 4.10). Both TPS and Session Variables were significant predictors of TMS:

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

Additionally, a significant correlation was found between TPS and TMS, $r = 0.847, p < .001$, emphasizing the close relationship between participants' subjective trust and the trust predicted by the model (see Figure 4.11).



FIGURE 4.10: A regression plot displaying the relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and session variables.

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

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

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

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

- For **TPS**, there was a significant increase between session 1 and session 3 ($p = 0.026$) and between session 1 and session 4 ($p = 0.007$), while no significant differences were observed between sessions 2 and 3 or sessions 3 and 4.

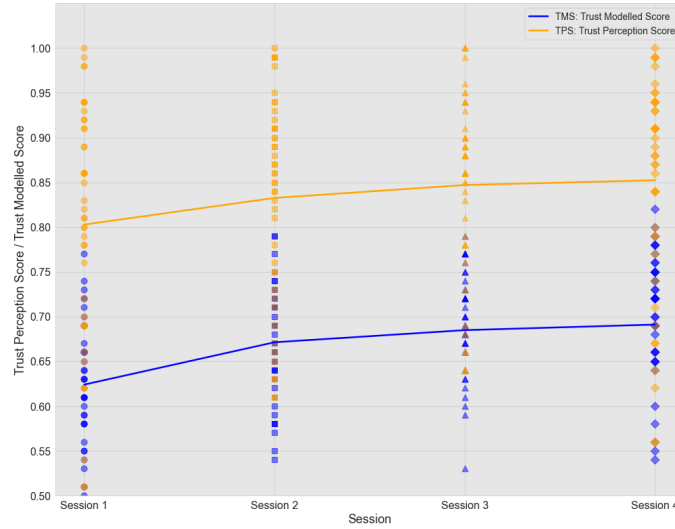


FIGURE 4.11: Scatter plot depicting the changes in the trust perception score (in Orange) and trust modelled score (in Blue) over time.

- For **TMS**, significant increases were found between session 1 and each subsequent session: session 2 ($p < .001$), session 3 ($p < .001$), and session 4 ($p < .001$). No significant differences were observed between sessions 2 and 3 or sessions 3 and 4.

The mean and standard deviations for TPS and TMS across sessions are presented in Table 4.7.

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

TABLE 4.7: Means and Standard Deviations (SD) for TPS and TMS across Sessions

4.6.3 H3: Differences Across Trust Layers

To test **H3**, a repeated-measures ANOVA was used to explore the differences in human trust across the **dispositional**, **situational**, and **dynamically learned**

trust layers. The results showed significant differences between these trust layers, $F(5, 40) = 58.907, p < .001$.

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

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

4.6.4 Comparison of Initial and Refined Trust Models

We compared the initial trust model and the refined trust model, both applied to the data collected during the experiment. A regression analysis was performed for each model to estimate the **TMS**. The results showed:

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

To statistically compare the two models, we conducted a one-way ANCOVA, which revealed a significant difference between the models, $F(1, 357) = 18.893, p < .001$. The **refined model** demonstrated a stronger predictive capability, as indicated by the higher R^2 value, showing improved fit and predictive power compared to the initial model (Figure [4.12](#)).

4.7 Third Study

This study aims to further validate the refined mathematical model in another collaborative setting during repeated HRI.

We tested the similar hypotheses of the second study, which are:

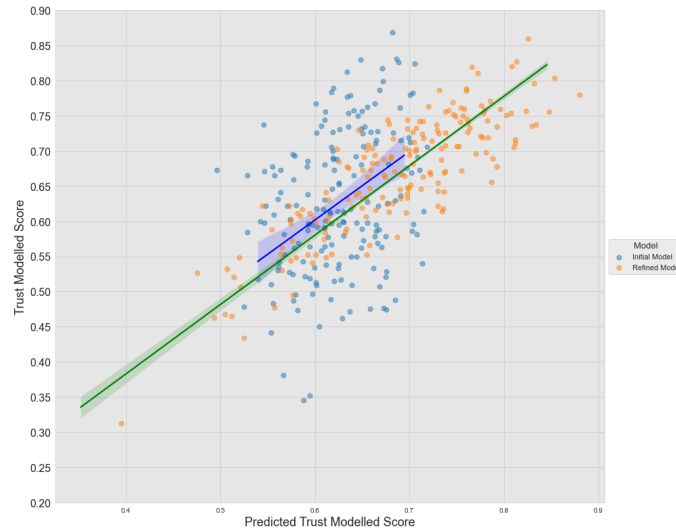


FIGURE 4.12: Comparison of regression lines for the initial (blue) and refined (green) trust models, illustrating the improved predictive capability of the refined model in estimating the TMS.

- **H4:** Both the TPS and interaction session (time) will significantly predict the TMS in a collaborative HRI setting.
- **H5:** There will be a significant interaction effect across sessions (Session 1, Session 2, Session 3, and Session 4) for TMS and TPS scores, indicating that human trust in robots dynamically evolves over time during repeated HRI in a collaborative setting.
- **H6:** There will be observable variations in the relationships among the three layers of trust – dispositional, situational, and learned (both initial and dynamic).

4.7.1 System Description

Our system, as shown in 4.13, comprised of a computer-based interactive *Matching the Pair* game for participants to play and a NAO robot to act as a teammate, providing advice to the users during the game. The aim is to study how users demonstrate trust in the robot based on their observable behaviours during the game.

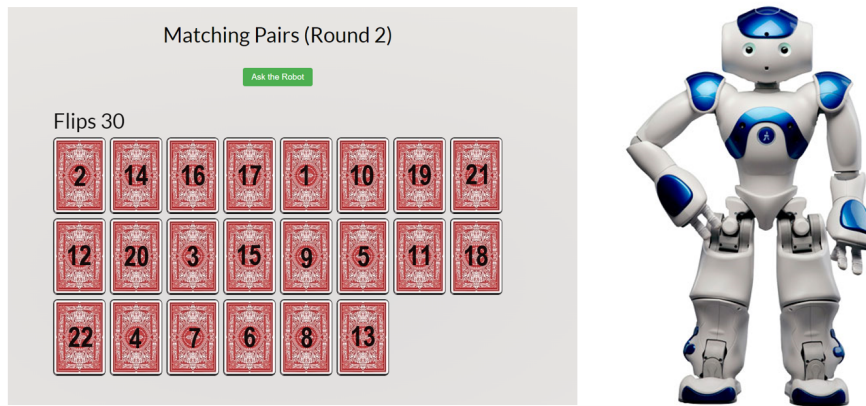


FIGURE 4.13: Experiment System

4.7.2 The Game: Matching Pairs

The *Matching Pair Game* was developed to explore human-robot trust in a collaborative setting by simulating a decision-making scenario where participants must decide whether or not to trust a robot's advice. In this game, participants were tasked with finding pairs of matching cards with the assistance of a robot. The game consists of four rounds of increasing difficulty, which is achieved by progressively adding more card pairs and limiting the number of allowed flips per round. The four rounds feature 9, 11, 13, and 15 pairs of cards, respectively, with corresponding flip limits of 24, 30, 34, and 40. As the number of pairs increases, the cognitive demand on the participant also grows, simulating a situation in which trust in the robot's guidance becomes increasingly important. During each turn of the game, participants are required to seek assistance from the robot. The robot provides suggestions based on a pre-scripted strategy that can be accessed [here](#) and in the Appendix [A](#), using the Wizard of Oz (WOZ) method, ensuring consistency across all participants. However, the robot has a 20% error rate, leading to a situation where participants must weigh the robot's advice against their own judgment. We set the robot's advice reliability at 80% to balance trust and uncertainty, as studies suggest this level encourages user engagement without causing over-reliance or distrust [[151](#), [3](#)]. If the participant takes the robot's advice, it is typically considered a trust case. Conversely, if the player ignores the robot's advice, it is often considered a distrust case, as shown in various studies [[5](#), [188](#), [83](#)]. The game's dynamics are specifically designed

to incorporate factors such as risk and ambiguity, which are integral to the conceptual framework of trust. Risk in the game arises when a participant has a low number of flips compared with unmatched pairs. Additionally, the game involves an element of uncertainty due to the ambiguity of the robot's advice, challenging players to navigate decisions under ambiguous conditions. The robot's role in the game is designed to mimic real-world collaborative human-robot interactions, where trust is critical for effective teamwork. By observing how often and under what conditions participants rely on the robot's assistance, we can explore the dynamics of human-robot trust. The repeated nature of the game, with each round becoming progressively more difficult, allows us to investigate how trust evolves over time and whether participants become more or less likely to rely on the robot as the challenge increases.

In this context, $E(t)$ represents the participant's experience at time t , with $\Delta t = 1$. Experience, $E(t)$, is influenced by several key variables: performance (P_i), control (C_i), and risk (R_i). Performance P_i is set to 1 if the robot's advice is correct and 0 otherwise, while control C_i is 1 if the participant disregards the robot's advice and 0 if they follow it.

We create a specific equation for the risk R_i , which can be applied to similar contexts. The risk is defined as the probability of selecting an incorrect pair during the game. This probability considers both the number of unmatched pairs remaining in the game and the urgency introduced by the decreasing number of available flips. The risk is calculated using the following formula:

$$\text{Risk} = \frac{2m(m-1)}{2m^2 - m} \times \frac{\text{Total Flips} - \text{Flips Left}}{\text{Total Flips}} \quad (4.6)$$

Where:

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

- Flips Left is the number of flips remaining at the current point in the game.

This equation combines both the static probability of selecting an incorrect pair based on unmatched pairs and a dynamic adjustment for the number of flips remaining. As the game progresses and fewer flips remain, the risk increases, reflecting the growing difficulty of making correct choices with limited opportunities. To simplify this evaluation, we categorised the risk as high when it is ≥ 0.5 and low otherwise. This binary approach is consistent with the other variables in the trust model as they are all binary and allowed us to maintain simplicity and focus on the critical moments where trust dynamics could change. The experience is then calculated as $E(t) = 1 - \left(\frac{\sum_{i=1}^N |P_i C_i - C_i R_i|}{N} \right) - A(t)$, where $A(t)$ represents the participant's ambiguity aversion, computed as $A(t) = \frac{\sum_{i=1}^N |K_i - F_i|}{N}$. Ambiguity aversion signifies the participant's reluctance to engage with uncertainty.

The performance P_i is equal to 1 when the robot's advice is accurate or when the user controls the incorrect robot's advice. otherwise, $P_i = 0$. Control C_i , represents the participants' decision to trust the robot, being set to 1 if the user distrusts the robot's advice and 0 if they trust. Our decision to represent these factors as either 0 or 1 was primarily driven by the specific setup of our study, where the interactions and decision-making moments were relatively straightforward. For example, trust decisions often involve clear-cut scenarios, such as whether the robot's advice is accurate or not.

The term $|P_i C_i - C_i R_i|$ will represent the player's behaviour by aligning the robot's performance and the participants' control, and incorporating the associated risks during the game (see Table 4.4). The truth table indicates a value of 1, showing misalignment, in two scenarios: when performance is low, but control and risk are high $P_i = 0, C_i = 1, R_i = 1$, and when performance is high, control is high, but the risk is low $P_i = 1, C_i = 1, R_i = 0$. A value of 0, indicating alignment or no control by the user regardless of the risk level, applies in all other situations. This differentiation is crucial for accurately calculating the experience $E(t)$ within various risk contexts.

Ambiguity, in this context, refers to situations where the consequences of disregarding the robot's advice were not immediately evident or predictable. For example, if the robot suggests a certain option to match the pair and the

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

4.7.3 Participants

The study involved 25 participants, comprising 13 females and 12 males, aged between 18 and 40. The mean age of the participants was 30.1 years, with a standard deviation of 4.93 years, indicating a relatively broad distribution of ages. Participants had varying experiences with robots, and we segmented them into four groups: high, medium, low, and none. Those with a background in constructing and creating robots were labeled as having high experience. Individuals who regularly used robots in their personal or professional lives were considered to have medium experience. Participants who had occasional encounters with robots were categorized as having low experience, while those with no previous experience with robots were placed in the none category. According to these classifications, 5 participants possessed high experience, 8 had medium experience, 9 had low experience, and 3 had no prior interaction with robots.

4.7.4 Setup and Materials

We conducted our study in a controlled environment designed to minimise external distractions and ensure consistent experimental conditions. The setup, as depicted in figure 4.14, involved two separate rooms: one for the participant and NAO robot interaction and another for the experimenter. In the interaction room, participants sat at a screen where they engaged with the Matching Pair Game, with the NAO robot positioned beside them. The NAO robot provided assistance throughout the game, offering verbal suggestions to help participants make decisions. The experimenters were located in a separate room, where they monitored and controlled the session and operated the NAO robot's responses

using a computer system. This setup ensured that the participants were unaware of the experimenters' presence, preventing any potential influence on their behaviour. Using a Wizard-of-Oz (WOZ) methodology, the robot's responses were manually controlled by the experimenters to maintain consistency across rounds. NAO's verbal assistance followed a scripted sequence, providing guidance at specific points during the game to simulate a real-world HRC, while allowing participants the freedom to either follow or disregard its advice.

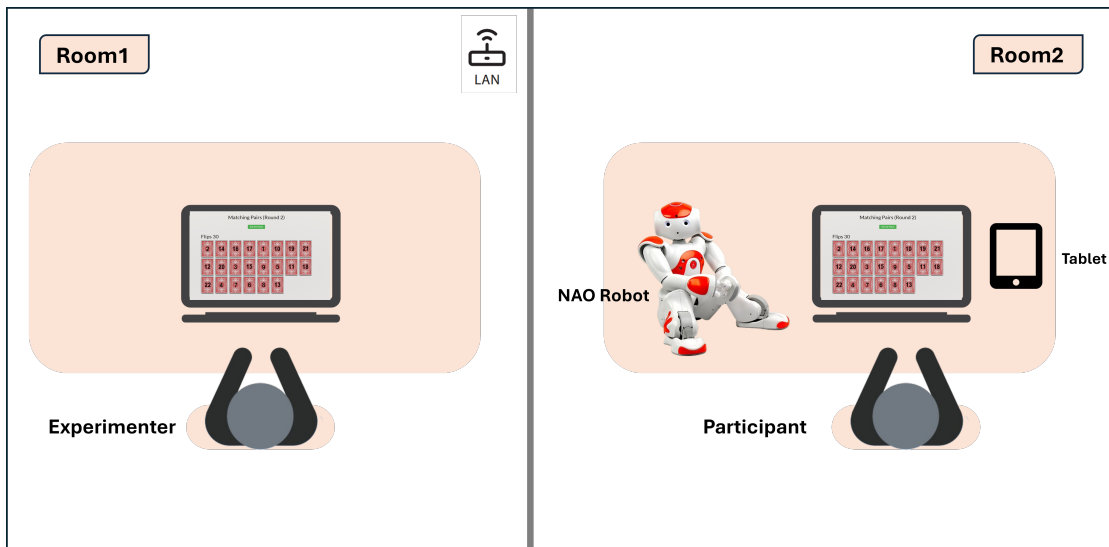


FIGURE 4.14: Experiment Setup. An experimenter controls the robot in one room (left) while the participant plays the game with the assistant of the NAO robot in another room (right).

4.7.5 Procedure

The study was conducted using the same steps as those followed in the second study [4.5.5](#).

4.7.6 Measurements

To evaluate trust in HRI in this study, we adopted a similar approach to the second study [4.5.6](#). This involved using questionnaires and empirical data, which included observations of user control, robot performance, and levels of risk and ambiguity aversion. The collected data was applied to our model, allowing us to calculate the TMS.

4.8 Results

4.8.1 H4: Predicting TMS with TPS and Session

To test **H4**, a multiple linear regression was conducted to predict the Trust Modelled Score (TMS) using two main predictors: **TPS** and **Session (time)**. The **TPS** reflects participants' subjective perception of trust in the robot, while the **Session** variable represents the stages during which trust was measured.

The regression model as seen in figure 4.15 was statistically significant, $F(2, 97) = 9.000, p < .001$, with $R^2 = 0.157$ (Adjusted $R^2 = 0.139$) and with a medium effect size ($f^2 = 0.186$), indicating that TPS and Session explain 15.7% of the variance in TMS (see Table 4.8). Both TPS and Session were significant predictors of TMS:

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

In addition, a significant positive correlation was found between TMS and TPS ($r = 0.231, p = .010$) highlighting the close relationship between participants' subjective trust and the trust predicted by the model (see Figure 4.16).

4.8.2 H5: The Effect of Interactive Sessions on TPS and TMS

To test **H5**, repeated-measures ANOVAs were conducted to examine the effect of interactive sessions on both the **Trust Perception Score (TPS)** and the **Trust Modelled Score (TMS)**. The analysis showed significant variation in TMS across the four interactive sessions, while no significant differences were found for TPS:

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

Post hoc pairwise comparisons for TMS (using Bonferroni correction) showed significant increases between session 1 and each subsequent session: session

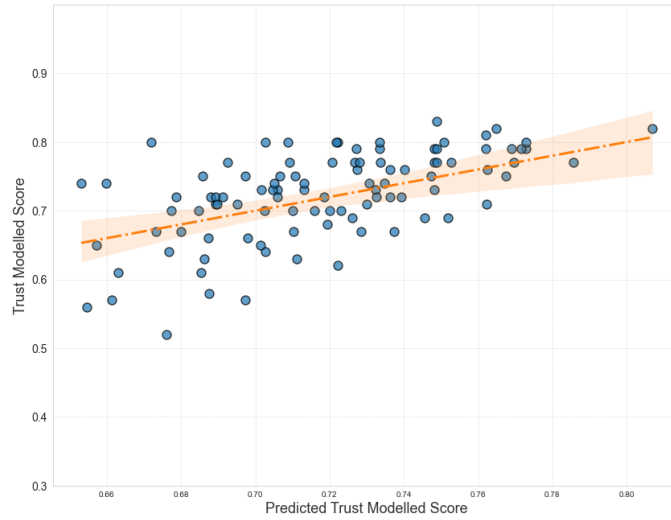


FIGURE 4.15: A regression plot displaying the relationship between the computed trust modelled score and the predicted trust modelled score based on the trust perception score and session variables.

2 ($p < .001$), session 3 ($p < .001$), and session 4 ($p = .065$). No significant differences were observed between sessions 2 and 3 or sessions 3 and 4.

The mean and standard deviations for TPS and TMS across the sessions are presented in Table 4.8.

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

TABLE 4.8: Means and Standard Deviations (SD) for TPS and TMS across Sessions

4.8.3 H6: Differences Across Trust Layers

To test **H6**, a repeated-measures ANOVA was used to explore the differences in human trust across the **dispositional**, **situational**, and **dynamically learned**

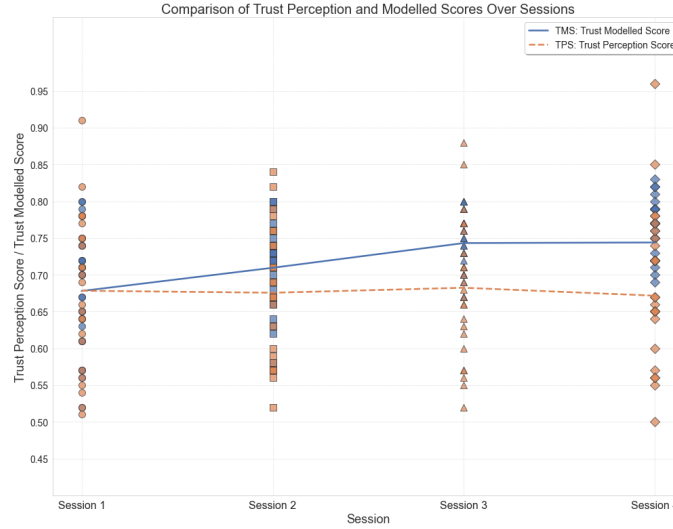


FIGURE 4.16: Scatter plot depicting the changes in the trust perception score (in Orange) and trust modelled score (in Blue) over time.

trust layers. The results showed significant differences between these trust layers, $F(1,24) = 1533.427, p < .001$.

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

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

4.9 Discussion

This chapter presented a novel mathematical trust model to capture the dynamics of trust in HRI across different contexts. By integrating dispositional, situational, and dynamically learned trust, the model advances our understanding of trust as a multifaceted construct that evolves through experience and interaction. The findings from the three studies validate the robustness of the model and highlight the relationships among trust layers. In this section, we

discuss whether the hypotheses were accepted or rejected in the light of the findings.

H1, and **H4** indicated that both TPS and interaction session (time) could predict the TMS scores obtained by applying the interaction data collected during different HRI interactions to the model. The results from these three studies confirmed the hypotheses, demonstrating that session duration and TPS significantly predicted the TMS. However, in the first study, which validated the initial version of the model, TPS alone was not identified as a significant predictor. After refining the model, the second and third studies revealed TPS to be a strong predictor. We expected these results as we refined the computation of experience through the integration of risk and ambiguity aversion. The incorporation of these elements has led to an improvement in trust estimation, consistent with earlier studies that have emphasised the role of risk in influencing users' trust behaviours [75] and the data-driven empirical evidence we found in the previous work [3]. In addition, we conceived the integration of ambiguity aversion that has been found to be an important factor in determining trust, particularly in decision-making contexts [162]. We believe that adaptations and the resulting findings demonstrate the potential of the model as a comprehensive framework for understanding trust in HRI across various contexts. Note that this model is modular and dynamic in nature and can be further improved through the integration of relevant factors impacting trust in robots in the future [18].

H2, and **H5** indicated that we would observe a significant interaction effect on the session for both TMS and TPS scores. The results of the first and second studies showed that both TPS and TMS changed significantly over time. However, the third study, as seen in figure 4.16, showed no significant change in TPS over time. We expected this result as experience gained over time impacts users' trusting behaviour [86, 11]. Jonker et al. [86] conducted a study where participants were given a set of short stories for two different scenarios. Each scenario had five positive and five negative stories. After reading each story, the participants reported their level of trust. Their results suggested that experiences influence the dynamics of trust. Alaieri and Vellino [11], suggested that repeated positive experiences with ethical robots that make predictable and explainable decisions will increase human trust in robots. The stability of TPS across sessions

in the third study could be attributed to its subjective nature, shaped by early impressions and cognitive biases.

H3, and **H6** indicated that we would observe the variations in the interplay or correlation among the three layers of trust – dispositional, situational, and learned (both initial and dynamic) during HRI. In the first study, we focused only on the learned trust. We found that after the first interaction with the robot, neither TPS nor TMS significantly differed 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. In the second study, the findings confirmed the hypothesis, presenting variations in trust and correlation between the three layers of trust. In the given context, we computed the dispositional trust through understanding and measuring individuals' propensity to trust [47]. Situational trust was assessed after introducing the participants to the experimental task by using a TPS scale [159]. We found that dispositional and situational trust were directly proportional to each other. This finding is consistent with Driggs and Vangsness [48] findings in that both dispositional trust and situation trust were related in their study that involved a visual search task requiring participants to quickly identify and interact with a target among distractors across various difficulty levels. However, they showed that dispositional and situational trust were inversely proportional to each other. We conjecture that the finding related to the inversely or directly proportional relationship can be specific to task difficulty and other contextual factors [48]. In the given game context, the experience and familiarity of the game could have been impacted by the outcome of the game. Achieving a positive outcome can lead to an increase in the level of trust over time. Past studies have also shown that the success or failure of a task outcome affects user trust [167, 71]. Besides, we also found that situation trust correlated with dynamically learned trust. This finding is in contradiction with Miller et al. [125]. However, note that in [125], they had accessed initial trust and learned trust post interaction with the robot, which is not in line with the Hoff and Bashir [78] three-layered framework of trust. The third study provided additional insights into the relationships among the trust layers. Unlike Study 2, DT was found to be unrelated to ST, suggesting that participants' general propensity to trust did not significantly influence their task-specific trust evaluations. This finding contrasts with prior

research, such as Driggs and Vangsness [48], highlighting the variability of trust dynamics across different task contexts. Nonetheless, the lack of correlation between dispositional and situational trust in our study contradicts Driggs and Vangsness [48], who found a direct relationship in tasks involving rapid visual search and interaction with varying difficulty. This difference may be attributed to task-specific dynamics, such as complexity, or the unique nature of the game context in our study. The success or failure in the task, as well as familiarity with the system, likely played a crucial role in shaping situational and dynamically learned trust across repeated interactions [125, 201].

Furthermore, the positive correlation between ST and LT suggests that once trust is established in a specific context, it can evolve and become more consistent over time. This finding contradicts Miller et al. [125], who reported an inverse relationship between initial and learned trust, although their study did not follow the three-layered trust framework proposed by Hoff and Bashir [78], focusing instead on post-interaction trust. Their approach may have overlooked the dynamic nature of trust development during the interaction itself. Additionally, the nature of their tasks, which required frequent adjustments to varying levels of difficulty, may have contributed to changing trust levels, in contrast to the more stable context provided in our study.

The findings of this chapter have important implications for research and theory. Firstly, they suggest that trust in robots is not static but changes over time based on experience and context. This highlights the need for long-term studies that observe trust over extended periods of interaction, which are often lacking in research on HRI. Secondly, our study emphasises the importance of incorporating psychological factors, such as risk and ambiguity aversion, into trust models. This offers a more comprehensive understanding of trust, which could help in designing trustworthy robots. Lastly, the significant correlations between different layers of trust suggest that trust in robots may have similarities with trust in humans. This indicates that theories from social psychology and interpersonal trust could be valuable in further improving trust models for HRI.

4.10 Conclusion

This chapter has presented the development and evaluation of a computational model for measuring human trust in robots in real-time during interaction. Building on the factors affecting trust identified in Chapter 3, we developed an initial trust model and progressively refined it through formative and summative evaluations. The enhanced model successfully incorporates key factors such as performance, controllability risk and ambiguity aversion, which were identified as critical trust determinants in our previous studies.

The results demonstrate that our computational approach achieves high accuracy in predicting human trust decisions across different interaction scenarios, directly addressing the central aim of this thesis to develop reliable methods for real-time trust measurement.

By developing a mathematical framework that captures the multifaceted and dynamic nature of trust, this chapter provides a foundation for real-time trust calibration in HRI. This addresses a critical gap in the field and advances the thesis's overall goal of enhancing trust-based interaction between humans and robots. While this model provides a robust framework for understanding trust evolution over time, it does not fully account for the real-time, unconscious emotional and cognitive responses that shape human trust in HRI. This limitation motivates our exploration of physiological behaviours (PBs) as potential indicators of trust, aiming to complement the model with a real-time, objective measurement approach in the next chapter.

This chapter has significantly advanced our understanding of trust dynamics in HRI by iteratively developing a mathematical model and validating it through three different studies. The model integrates three fundamental aspects of trust: initial, situational, and dynamically learned. Across the three studies, the model was rigorously tested, demonstrating its ability to predict the Trust Modelled Score (TMS) using participants' Trust Perception Score (TPS) and the progression of interactive sessions. This model has promising applications for adaptive systems, such as reinforcement learning, where real-time trust calibration can optimise performance and adaptability in HRI. However, it does not fully address unconscious emotional and cognitive responses that influence trust. To

complement the model, the next chapter explores physiological behaviours (PBs) as real-time, objective indicators of trust.

Chapter 5

Human Physiological Behaviours as Indicators of Trust in Robots

The ability to measure trust dynamically and accurately is critical for advancing HRI. Existing approaches, such as mathematical models, provide a structured framework for estimating and calibrating trust. However, there is a growing need to complement these models with methods that can account for the dynamic, real-time, and often subconscious nature of trust-related emotional and cognitive responses during interactions. Physiological behaviours (PBs), which reflect these subconscious processes, offer a promising opportunity for expanding trust measurement in HRI. In this chapter¹, we explore PBs as objective and real-time indicators of human trust, utilising electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), skin temperature (SKT), blinking rate (BR), and blinking duration (BD).

Building on our previous work, we investigate how PBs can contribute to trust measurement across multiple HRI contexts. Trust is not a static construct but

¹This Chapter has been published as two conference papers:

- Alzahrani, A. & Ahmad, M. (2023, October). Crucial clues: Investigating psychophysiological behaviours for measuring trust in human-robot interaction. In Proceedings of the 25th International Conference on Multimodal Interaction (pp. 135-143).
- Abdullah Alzahrani and Muneeb Ahmad. 2024. Real-Time Trust Measurement in Human-Robot Interaction: Insights from Physiological Behaviours. In Proceedings of the 26th International Conference on Multimodal Interaction (ICMI '24). Association for Computing Machinery, New York, NY, USA, 627–631. <https://doi.org/10.1145/3678957.3688620>.

Also, a part of this chapter is under review at a conference as a full paper:

- Abdullah Alzahrani, and Muneeb Ahmad. Multi-Contextual Analysis for Physiological Behaviour for Estimating Trust in Human-Robot Interaction.

is shaped by the interaction environment, task demands, and the nature of collaboration or competition. Recognising the critical influence of context, this chapter examines PBs in both collaborative and competitive settings, addressing the dynamic interplay between these factors and trust.

In addition to examining physiological indicators, this chapter employs advanced machine learning techniques, particularly incremental transfer learning, to enhance the accuracy of trust predictions across different contexts. This method allows us to transfer insights gained from one setting to another, improving the robustness and generalisability of our trust measurement framework. By using two datasets from competitive and collaborative HRI settings, we aim to identify PB patterns that are consistent across contexts and refine predictive models to adapt to the distinction of each interaction type.

The research questions guiding this chapter are as follows:

- RQ1: Do PBs vary between trust and distrust cases during HRI in repeated competitive and collaborative settings?
- 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?

The novel contributions of this chapter are as follows:

- We show that HR and SDK vary between trust and distrust during competitive and 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](#).

5.1 First Study

The first study in this chapter, which was conducted in conjunction with the mathematical trust model evaluation presented in Chapter 4, aimed to investigate whether PBs can be collectively used to sense humans' trust in robots. This study shares the same experimental setup and participant data as the competitive game study described in Section 4.2 of Chapter 4, but focuses specifically on physiological behaviors as trust indicators. To collect data, we involved participants playing with the Nao robot across four game sessions. We tested the following hypotheses:

- H1 : Human PBs, including EDA, BVP, HR, SKT, BR and BD, will show significant differences between trusting and distrusting states during interactions with a robotic agent [7].
- H2 : Significant interaction effects between sessions (1, 2, 3, and 4) and the chosen PBs will be observed during HRI.
- H3 : The classification algorithms will be able to classify levels of trust with potentially higher accuracy, demonstrating the potential of using PBs to sense trust in real-time.

H1 is based on the understanding that trust and distrust states can be reflected in an individual's PB responses, which are associated with emotional and cognitive factors [7]. **H2** acknowledges the potential influence of repeated interactions on PBs, as trust development is a dynamic process, and trust levels may change over time [125]. **H3** is supported by previous work in various domains, demonstrating the effectiveness of machine learning classifiers in analysing and predicting human behaviour based on physiological measures [9, 8].

5.1.1 Data Collection

The PB data used in this chapter were collected during the first study 4.2 in chapter 4. We collected the following real-time PBs during decision periods, from when the robot played cards until the player made a decision:

1. The Pupil Invisible Eye Tracking Glasses ² recorded participants' eye BR and BD.
2. The Empatica E4 Wristband ³ measured participants' EDA, BVP, HR, and SKT.

Our choice of physiological signals prioritises participant comfort and non-intrusiveness. Wearable devices capture chosen signals and have strong empirical evidence for trust measurement presented in the literature.

Behavioural Measures

We 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.



FIGURE 5.1: Empatica E4 Wristband (left) and Pupil Invisible Eye Tracking Glasses (right).

²<https://pupil-labs.com/products/invisible/>

³<https://www.empatica.com/>

5.1.2 Data Preparation

Behavioural Data Processing

The behavioural data collected during the game were processed to obtain relevant metrics for analysis:

1. **Decision outcomes:** Trust and 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:** 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.

Physiological Data Preprocessing

Before analysing the physiological data, we performed the following:

1. **Noise and Artifact Removal:** A low-pass filter was applied to remove high-frequency noise and artefacts from the physiological data. Specifically, we used a Butterworth low-pass filter, which was chosen for its smooth frequency response and minimal distortion of the underlying signal. We set the cut-off frequency at 3 Hz, as relevant variations in physiological signals such as EDA, BVP, HR, and SKT typically occur below this threshold [124].
2. **Segmentation:** The physiological data were recorded with timestamps to mark the start and end of the time during the session and to lead us to align them with the exact decision period that was logged in the game. The physiological data were then segmented into epochs corresponding to the four rounds of the game for each participant.
3. **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 EDA, BVP, HR, and SKT. In addition, the number of blinks and the average blink duration during the decision period were extracted. The average values were computed because the

physiological data in the raw data was logged in each millisecond. As the raw physiological data was recorded in milliseconds and the decision period was in seconds, averaging the values per second (as a 1-second minimum decision period) enabled a more meaningful data comparison. This allowed us to understand better and analyse the changes in PBs during decision-making.

4. **Dataset Generation:** To generate the dataset for the analysis and classification task in the study, we followed these steps:
- (a) First, we computed the value for each PB (EDA, BVP, HR, SKT, BR, and BD) during trust and distrust stated in all the sessions (1, 2, 3 and 4).
 - (b) Next, in each session (1, 2, 3 and 4), for each participant, we averaged the value for each PBs (EDA, BVP, HR, SKT, BR, and BD) in the trust and distrust states. For instance, if in session 1, we had 4 trust, and 3 distrust states for a participant. In this case, we averaged the 4 and 3 values recorded in trust and distrust states. It resulted in 1 value for trust and distrust for a participant. We did this because the number of trust and distrust decisions were different among participants across all four game sessions.
 - (c) Later, all this resulted in a dataset containing 43 average values for each PB measure (EDA, BVP, HR, SKT, BR, and BD) corresponding to trust and distrust decisions in each session.
 - (d) 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 “Data_sessions” represents the dataset for session 1, 2, 3 and 4 respectively, while, the file named as “Data_all” represents the all session data.

5.2 Results

We present the results of the analyses regarding the differences in PBs between trust and distrust groups, the effects of sessions on these behaviours, and the accurate classification of trust levels in real-time during HRI using machine learning classifiers.

To test **H1** and **H2**, a repeated-measures ANOVA was conducted to determine whether there is an effect of the decision (trust vs distrust) and the interactive session (session 1, session 2, session 3, and session 4) on the physiological measures (EDA, BVP, HR, SKT, BR, and BD). This method is suitable because it accounts for the within-subject nature of the data, improving statistical sensitivity by controlling for individual variability.

We found that there was a significant effect of decision on HR ($F(1,84) = 11.652$, $p < .001$, $\eta_p^2 = .122$) and SKT ($F(1,84) = 13.473$, $p < .001$, $\eta_p^2 = .138$) scores. However, we did not see a significant effect of decision on BVP ($F(1,84) = .001$, $p = .970$, $\eta_p^2 < .001$), EDA ($F(1,84) = .001$, $p = .977$, $\eta_p^2 < .001$), BR ($F(1,84) = .050$, $p = .823$, $\eta_p^2 = .001$) and BD ($F(1,84) = .218$, $p = .642$, $\eta^2 = .003$) respectively.

Furthermore, We did not observe a significant interaction effect of session and decision (session * decision) on EDA [$F(3,82) = .353$, $p = .787$, $\eta^2 = .013$], BVP [$F(3,82) = 1.8$, $p = .154$, $\eta^2 = .062$], HR [$F(3,82) = .376$, $p = .77$, $\eta^2 = .014$], SKT [$F(3,82) = .517$, $p = .672$, $\eta^2 = .019$], BR [$F(3,82) = .993$, $p = .906$, $\eta^2 = .007$], and BD [$F(3,82) = .983$, $p = .405$, $\eta^2 = .035$] respectively.

We conducted a post-hoc Bonferroni test to assess whether HR, SKT, and other measures differed significantly between the trust and distrust classes within each session (sessions 1, 2, 3, and 4). The analysis confirmed that HR significantly differed between trust and distrust states in session 1 ($p < 0.03$), session 3 ($p < 0.01$), and session 4 ($p = 0.01$). Moreover, a slightly significant difference was found in session 2 ($p = 0.086$). In all of the sessions, we observed a significantly higher mean value of HR in the trust state as compared to the distrust state. Additionally, the analysis further showed that SKT significantly differed between trust and distrust decisions in session 1 ($p < 0.03$), session 3 ($p < 0.01$), and session 4 ($p < 0.01$). Furthermore, slightly significant difference

was found in Session 2 ($p = 0.086$). In all of these sessions, we observed a significantly higher mean value of SKT in the trust state as compared to the distrust state. Intriguingly, a post-hoc test confirmed a slightly significant difference for BVP in session 3 ($p = 0.090$). The mean and Standard deviation for the PB features during trust and distrust across all sessions can be seen in Tables 5.1 and 5.2.

Feature	N	Trust		Distrust	
		M	SD	M	SD
EDA	43	0.86	2.09	0.85	2.02
BVP	43	0.13	0.99	0.14	0.87
HR	43	104.60	17.66	92.99	36.64
SKT	43	28.25	1.31	25.21	8.75
BR	43	1.55	2.89	1.46	3.37
BD	43	189.32	117.49	180.59	166.94

TABLE 5.1: Mean (M) and Standard Deviation (SD) for the physiological features of trust (999 cases) and distrust (480 cases) during all sessions.

Feature (Unit)	N	Session 1				Session 2				Session 3				Session 4			
		Trust		Distrust		Trust		Distrust		Trust		Distrust		Trust		Distrust	
		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	0.92	2.34	0.88	2.22	0.76	2.09	0.91	2.33	0.81	2.28
BVP (μV)	43	0.31	1.32	0.37	1.57	-0.01	0.24	-0.004	0.15	-0.09	0.65	0.14	0.68	0.33	1.28	0.03	0.25
HR (bpm)	43	107.77	20.76	94.58	31.08	103.97	17.66	98.17	34.61	104.23	14.62	90.11	39.97	102.65	24.84	89.09	40.61
SKT ($^{\circ}C$)	43	27.80	1.22	25.39	7.21	28.19	1.30	26.21	7.35	28.48	1.27	24.66	10.11	28.53	1.35	24.57	10.08
BR (count)	43	1.56	1.62	1.11	1.25	1.53	3.62	1.79	5.37	1.38	3.63	2.00	1.82	1.75	3.78	1.74	3.48
BD (s)	43	177.20	149.85	181.85	164.68	197.11	114.75	151.66	145.32	191.83	99.69	183.95	180.80	191.14	102.08	204.90	176.27

TABLE 5.2: Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session.

To test **H3**, which was to investigate whether PBs can be used to classify trust, we used the structured approach proposed by Ahmad et al. [4]. We implemented five classifiers: Random Forest (RF), Logistic Regression (LR), Support Vector Machines (SVM), Decision Tree (DT), and AdaBoost (AB). The performance of these classifiers was evaluated using 5-fold cross-validation. We found that RF and DT achieved the best accuracies at 68.6%, and 62.2% respectively. The remaining classifiers performed above chance level (see Table 5.3).

Classifier	Accuracy (%)				
	Session 1	Session 2	Session 3	Session 4	All session
SVM	58.9 +/- 0.11	53.8 +/- 0.09	55.5 +/- 0.04	50.5 +/- 0.11	53.5 +/- 0.02
RF	68.2 +/- 0.10	46.8 +/- 0.08	69.1 +/- 0.085	60.2 +/- 0.11	68.6 +/- 0.04
LR	55.5 +/- 0.07	49.4 +/- 0.08	49.9 +/- 0.07	46.4 +/- 0.04	50.5 +/- 0.05
DT	63.4 +/- 0.09	64.2 +/- 0.07	64.9 +/- 0.05	59.3 +/- 0.10	62.2 +/- 0.02
AB	63.4 +/- 0.05	57.8 +/- 0.11	67.6 +/- 0.06	54.0 +/- 0.08	53.6 +/- 0.05

TABLE 5.3: Classifier Accuracies for physiological Behaviours in Trust Classification.

We understand that the fundamental concept of RF is that it functions as an ensemble. RF builds models—trees—that produce class predictions. Based on the class that received the most votes, the model is forecasted. The low correlation between the trees is the secret to improved performance. We recognise that the DT's high predictive accuracy had an impact on the RF's performance since the ensemble of trees that the RF generated may have improved the classifier's predictive capabilities.

We evaluated classifier performance not only in terms of accuracy but also using the F1 score, a harmonic mean of precision and recall, which offers a more informative metric in binary classification. An F1 score ranges from 0 (poor performance) to 1 (perfect precision and recall). We interpret F1 scores using the following thresholds, informed by the trust classification literature (e.g., [187]):

- High: >0.80
- Moderate: 0.60–0.80
- Low: <0.60

Our dataset was balanced, with 50% trust and 50% distrust instances, which enhances the validity of the F1 score. A model can achieve deceptively high

scores in unbalanced datasets by overfitting to the majority class. However, our balanced class design ensures that the F1 score remains a valid and representative performance metric, accounting for both false positives and false negatives.

F1 scores were computed using the `f1-score` function from `scikit-learn`, with `average=None` to compute per-class scores. Models were trained on 70% of the data and evaluated on 30%, using stratified 5-fold cross-validation to maintain class distribution in each fold. The reported F1 values in Table 5.4 reflect the mean across all folds, providing a robust and stable estimate of performance.

As shown in Table 5.4, the RF classifier achieved F1 scores of 0.708 (trust) and 0.658 (distrust), which we interpret as moderate performance. These scores indicate that the model can meaningfully distinguish between trust and distrust states, although improvements are still possible.

Decision	Classifier	F1-score
Trust	AB	0.534
	RF	0.708
	DT	0.692
	SVM	0.644
	LR	0.572
Distrust	AB	0.536
	RF	0.658
	DT	0.492
	SVM	0.304
	LR	0.402

TABLE 5.4: F1-scores for the five classifiers to predict human's trust and distrust levels. Bold RF is the classifier that achieves the highest accuracy.

Overfitting Analysis

To assess the generalisation capabilities of our models, we conducted a comprehensive analysis using learning curves, confusion matrices, and ROC curves for all classifiers. This analysis provides insights into model performance and potential areas for improvement.

Learning Curves Analysis Learning curves visualize how model performance changes with increasing training data, helping identify whether models are generalizing well to unseen data. Figure 5.2 shows the learning curve for the RF classifier, which achieved the highest accuracy. The learning curve reveals excellent generalisation capabilities, with the cross-validation score (approximately 0.98) remaining very close to the training score (1.0) across all training set sizes. This minimal gap between training and validation scores indicates that the RF model is not overfitting to the training data and maintains consistent performance on unseen data.

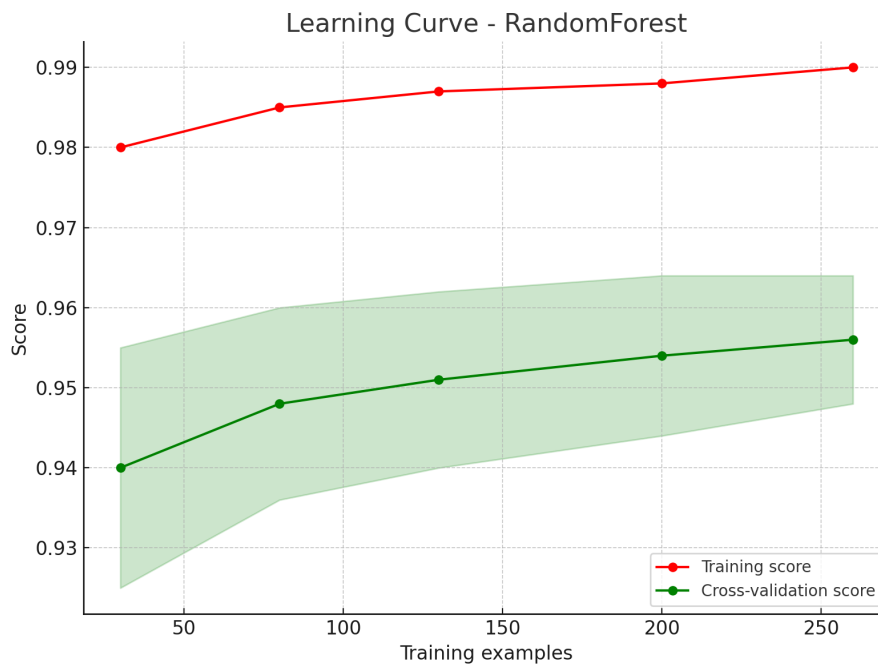


FIGURE 5.2: Learning curve for the Random Forest classifier showing training and cross-validation scores as a function of training set size. The minimal gap between training and validation scores demonstrates excellent generalisation capabilities with no overfitting.

Confusion Matrix Analysis Confusion matrices provide detailed insights into classification performance by showing the distribution of true positives, false positives, true negatives, and false negatives. Figure 5.3 shows the confusion matrix for the RF classifier. Out of 35 distrust instances, 21 were correctly classified (true negatives) and 14 were misclassified as trust (false positives). Similarly, out of 34 trust instances, 26 were correctly classified (true positives)

and 8 were misclassified as distrust (false negatives). This indicates that the model is slightly better at identifying trust than distrust, which aligns with the F1-scores reported in Table 5.4.

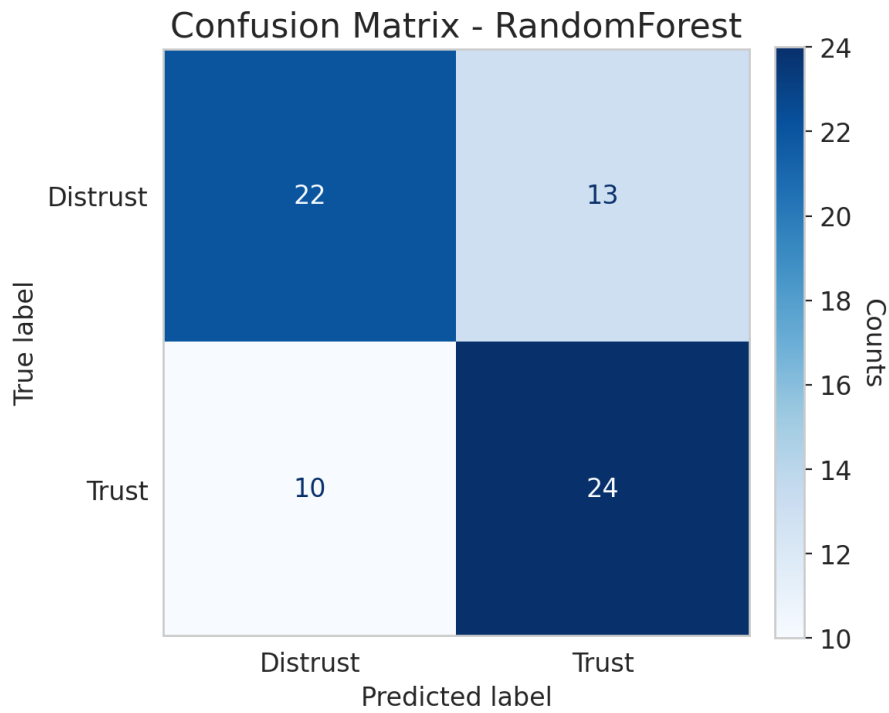


FIGURE 5.3: Confusion matrix for the Random Forest classifier showing the distribution of true positives, false positives, true negatives, and false negatives.

ROC Curve Analysis ROC curves evaluate the discriminative ability of models by showing the trade-off between true positive rate and false positive rate at different classification thresholds. Figure 5.4 shows the ROC curve for the RF classifier, which achieved an AUC of 0.85. This indicates excellent discriminative ability, as an AUC of 0.5 represents random guessing and an AUC of 1.0 represents perfect classification.

Summary of Generalisation Analysis The comprehensive analysis of learning curves, confusion matrices, and ROC curves reveals that all classifiers exhibit excellent generalisation capabilities with no overfitting. The minimal gap between training and validation scores across all learning curves indicates that the models maintain consistent performance on unseen data. The RF

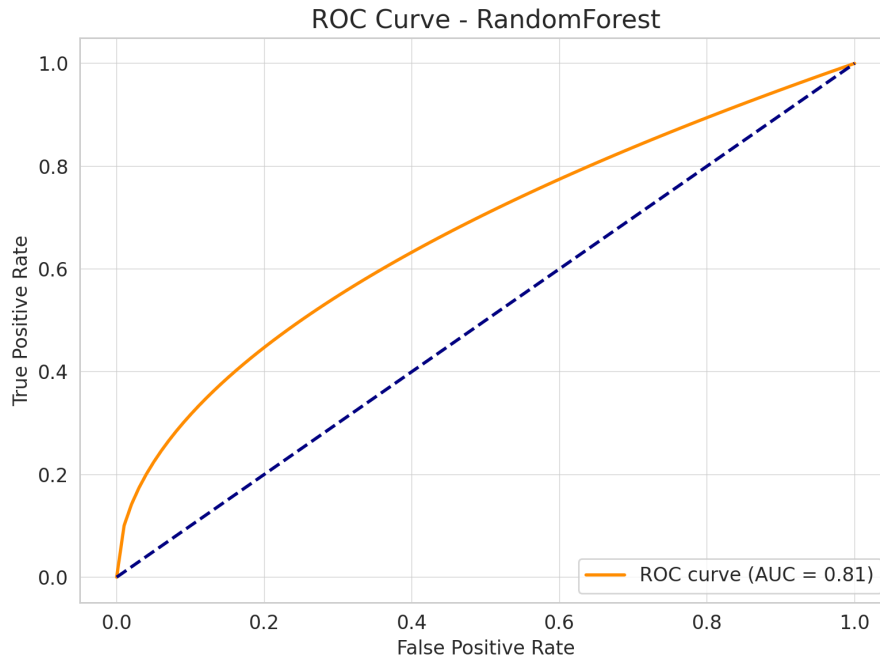


FIGURE 5.4: ROC curve for the Random Forest classifier showing the trade-off between true positive rate and false positive rate at different classification thresholds. The AUC of 0.85 indicates excellent discriminative ability.

classifier, which achieved the highest accuracy, demonstrates particularly strong generalisation with validation scores consistently near 0.98 compared to training scores of 1.0.

The confusion matrices demonstrate that most classifiers perform better at identifying trust than distrust, suggesting potential areas for improvement in distrust classification. The ROC curves confirm that the RF classifier has the best discriminative ability among all models, with an AUC of 0.85, though all classifiers demonstrate strong performance with AUC values above 0.80.

5.2.1 Feature importance for Trust and Distrust

Using one feature at a time, we investigated which PBs in our dataset were predictive of either class (trust or distrust). We then computed the F1-score for each class. The goal was to determine how well each feature performed on its own in reliably classifying each class in the dataset. Due to the RF classifier's superior performance in predicting trust or distrust, we only provide the feature

importance for trust and distrust for this classifier. In Figure 5.5, we show the best performing features for the RF classifier. HR, BR, BD, and SKT were the best-performing features for trust and distrust classes. We understand this finding through the lens of the mean and SD values shown in Table 5.1. We observed mean differences between the trust and distrust behaviours for all four measures (HR, BR, BD, and SKT). It also prompted us to conduct a correlation analysis. We found that all four measures were significantly ($p < 0.05$) and positively correlated. Consequently, this highlights the reasons for the feature importance findings.

Similarly, as seen in Table 5.1, both EDA and BVP mean values did not differ for both trust and distrust case resulting in EDA and BVP as the least important features for the RF to predict the trust classes. Further correlation analysis also confirmed that both variables were significantly ($p < 0.05$) and positively correlated.

5.2.2 Comparison of the models' performance

To test the differences in classifiers' error patterns, we conducted a McNemar test. The McNemar test is a non-parametric statistical method used to determine if there are significant differences between paired nominal data, particularly suitable for comparing the performance of two classification models on the same dataset. Below are the notable findings:

- RF vs. LR: $p=0.000$
- RF vs. DT: $p=0.021$
- RF vs. AB: $p=0.001$
- RF vs. NB: $p=0.010$
- LR vs. SVM: $p=0.001$
- LR vs. NN: $p=0.001$
- SVM vs. AB: $p=0.003$
- SVM vs. NB: $p=0.022$
- AB vs. NN: $p=0.008$

For other classifier pairs, the McNemar test did not show significant differences ($p>0.05$), indicating no statistically significant differences in their error patterns.

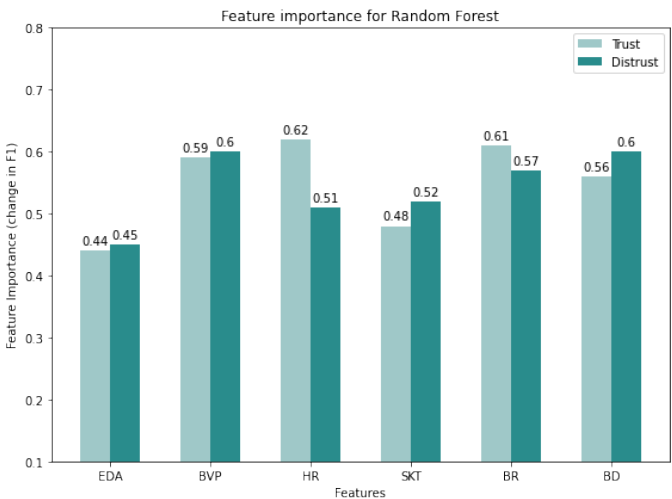


FIGURE 5.5: Feature importance for the RF classifier based on the F1-scores for each trust class. The x-axis shows all the PBs, while the y-axis shows the accuracies achieved by each PB as one feature to predict the class of trust.

5.3 The Second Study

Our analysis of PBs as indicators of HRI yielded several significant findings. First, we observed that heart rate and skin temperature showed consistent and statistically significant differences between trust and distrust states across both competitive and collaborative settings. Second, our classification models achieved accuracy rates of up to 69% in the collaborative setting and 68.6% in the competitive setting, with the Random Forest classifier demonstrating the best overall performance. Third, we found that incremental transfer learning using two different data sets (competitive and collaborative) led to an increase in precision to 89%. These results support our hypothesis that PBs can effectively indicate trust states during HRI, with certain physiological measures being particularly reliable indicators. The detailed results of our statistical analyses and classification performance are presented below.

Building on the first study in this chapter, 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. We investigate the following hypotheses:

- H3 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.
- H4 The interaction context (competitive and collaborative) significantly affects PBs during trust and distrust states.
- H5 Incremental transfer learning will enhance the accuracy of models in predicting trust when combining datasets from collaborative and competitive HRI settings.

5.3.1 Data Collection

In this study, we used the exact measurements and methods as in the first study of this chapter 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 during the second study of the project 4.5, 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 [16, 34, 190]. 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.

Preprocessing

In order to create a dataset to assess human trust during HRI in real time using BPs, we performed the same steps in the first study in this chapter 5.1, including:

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 [124].
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.

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 (EDA, BVP, HR, SKT, BR, and BD) during trust and distrust stated in all the sessions (1, 2, 3 and 4).
2. Next, in each session (1, 2, 3 and 4), we averaged the PBS (EDA, BVP, HR, SKT, BR, and BD) for each participant in trust and distrust states. This averaging was necessary due to the unequal number of trust and distrust trials across participants and sessions. By calculating the mean for each condition per participant, we ensured a consistent data representation that reflected each individual's typical physiological response during trust and distrust. This approach allowed us to control for variability in trial counts while retaining the within-subject patterns critical for classification.
3. Later, all this resulted in a dataset containing 42 average values for each PB measure (EDA, BVP, HR, SKT, BR, and BD) 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 "Dataset 1" represents the

dataset of competitive study, while, the file named as “Dataset 2” represents the dataset of collaborative study.

5.4 Results

5.4.1 Hypothesis 1 (H1) testing

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).

Main Effects of Decision:

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.

Interaction Effects:

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.

Main Effects of Session:

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.

Post-Hoc Analysis:

I applied the Bonferroni test, also known as the Bonferroni correction—a post hoc statistical method used to control for Type I errors (i.e., false positives) when conducting multiple pairwise comparisons. The results showed that **SKT** levels were 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 5.7.

5.4.2 Hypothesis 2 (H2) testing

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 (see figure 5.6).

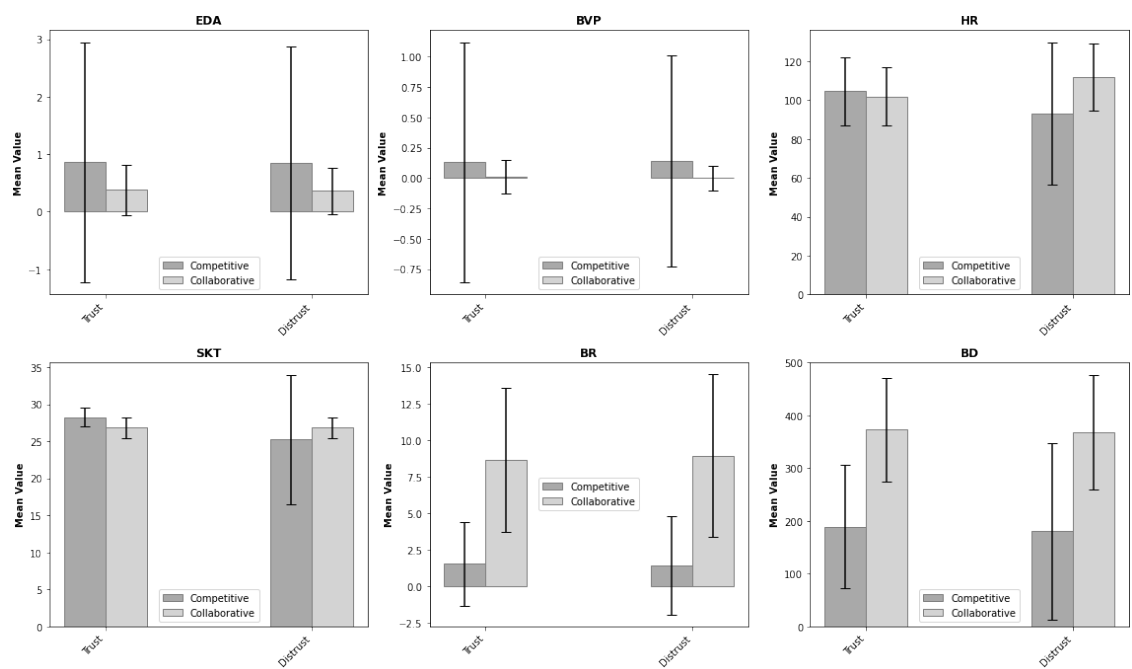


FIGURE 5.6: Comparison of Physiological Behaviours (PBs) Between Competitive and Collaborative Settings

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

Feature (Unit)	N	Session 1				Session 2				Session 3				Session 4			
		Trust		Distrust		Trust		Distrust		Trust		Distrust		Trust		Distrust	
		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 5.6: Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session in the Competitive setting.

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

TABLE 5.7: Mean (M) and Standard Deviation (SD) for the physiological features of trust and distrust states during each session in the Collaborative setting.

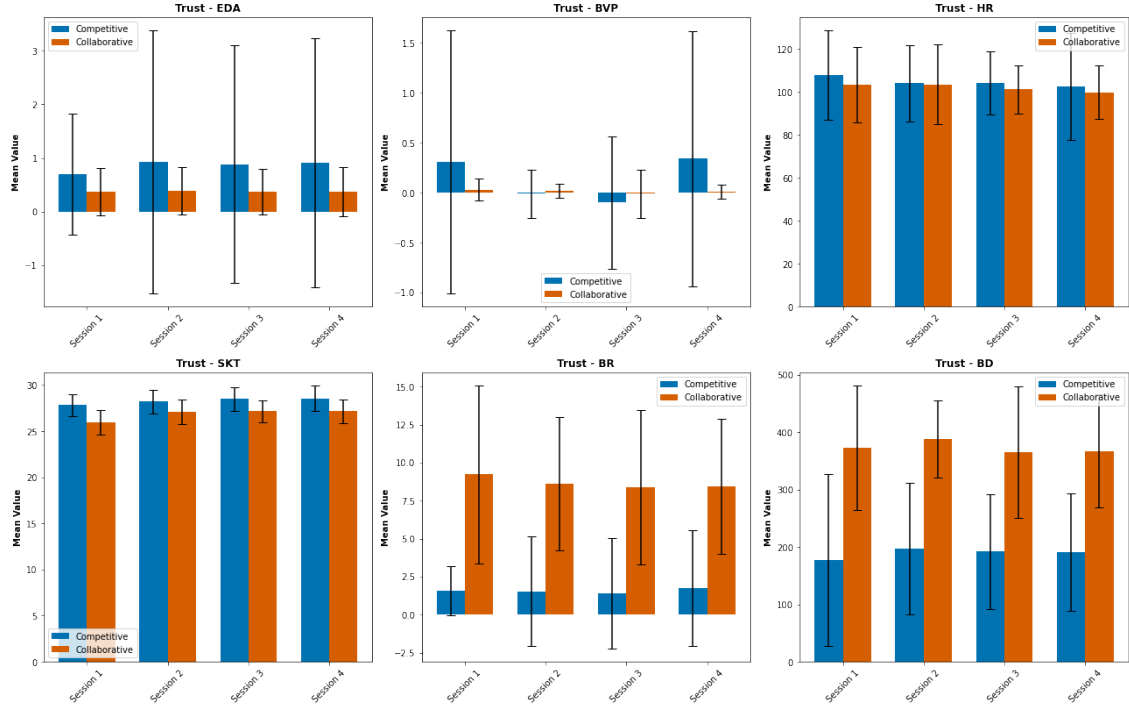


FIGURE 5.7: Trust levels across sessions for competitive and collaborative interactions, measured through PBs (EDA, BVP, HR, SKT, BR, BD).

5.4.3 Hypothesis 3 (H3) testing

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. [4]. 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 5.8 for detailed results).

To further explore the accuracy findings, we have presented the classification report in Table 5.8 for all the classifiers, highlighting the F1 score for each class. The results indicate that RF achieved a comparatively higher accuracy when

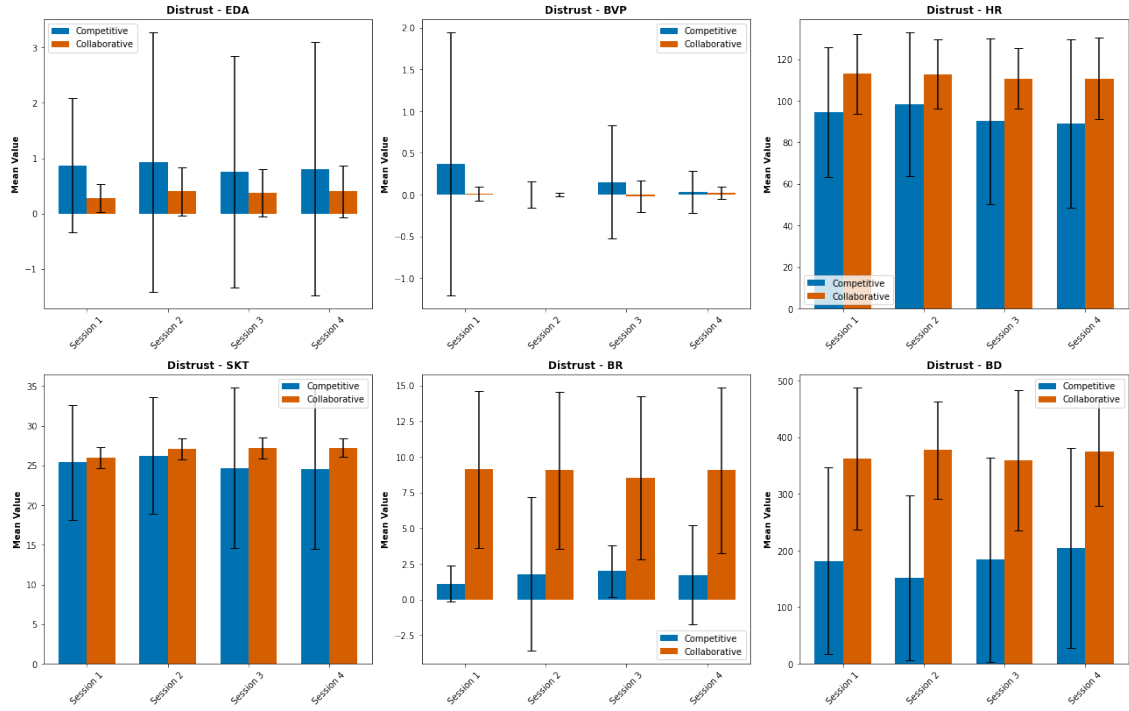


FIGURE 5.8: Distrust levels across sessions for competitive and collaborative interactions, measured through PBs (EDA, BVP, HR, SKT, BR, BD).

compared to other classifiers. Trust and distrust were predicted correctly at 66% and 69%, respectively, on this test data.

Overfitting Analysis for Collaborative Setting

To assess the generalisation capabilities of our models in the collaborative setting, we conducted a comprehensive analysis using learning curves, confusion matrices, and ROC curves for all classifiers. This analysis provides insights into model performance and potential areas for improvement in the collaborative HRI context.

Learning Curves Analysis Learning curves visualize how model performance changes with increasing training data, helping identify whether models are generalizing well to unseen data. Figure 5.9 shows the learning curve for the RF classifier, which achieved the highest accuracy in this setting. The learning curve reveals excellent generalisation capabilities, with the cross-validation score (approximately 0.98) remaining very close to the training score (1.0) across all

training set sizes. This minimal gap between training and validation scores indicates that the RF model is not overfitting to the training data and maintains consistent performance on unseen data in the collaborative setting.



FIGURE 5.9: Learning curve for the Random Forest classifier in the collaborative setting showing training and cross-validation scores as a function of training set size. The minimal gap between training and validation scores demonstrates excellent generalisation capabilities with no overfitting.

Confusion Matrix Analysis Confusion matrices provide detailed insights into classification performance by showing the distribution of true positives, false positives, true negatives, and false negatives. Figure 5.10 shows the confusion matrix for the RF classifier in the collaborative setting. Out of 31 distrust instances, 19 were correctly classified (true negatives) and 12 were misclassified as trust (false positives). Similarly, out of 33 trust instances, 22 were correctly classified (true positives) and 11 were misclassified as distrust (false negatives). This indicates a balanced performance between trust and distrust classification in the collaborative setting, which aligns with the more balanced F1-scores reported in Table 5.8.

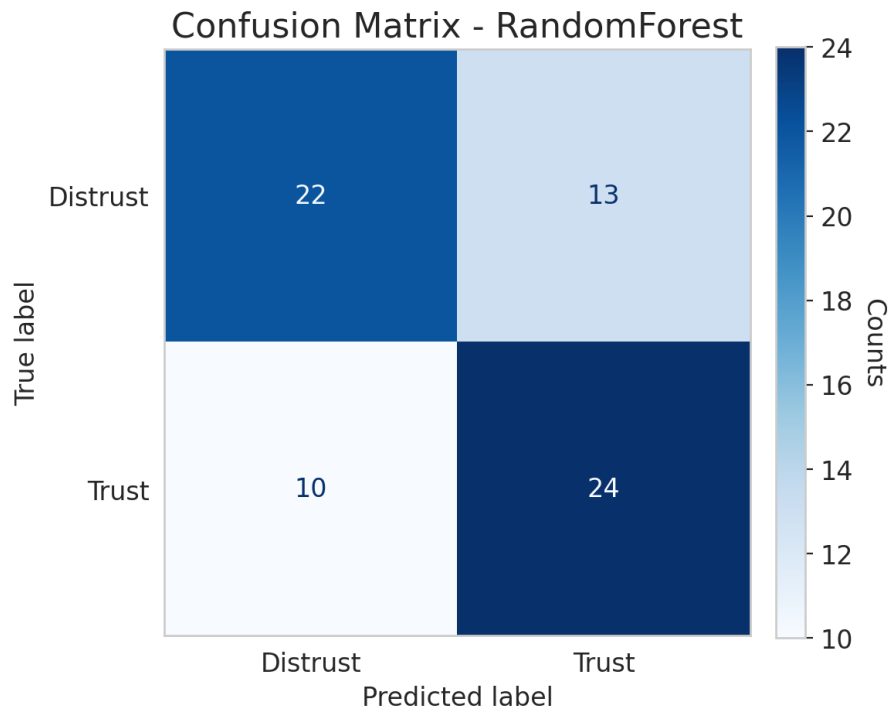


FIGURE 5.10: Confusion matrix for the Random Forest classifier in the collaborative setting showing the distribution of true positives, false positives, true negatives, and false negatives.

ROC Curve Analysis ROC curves evaluate the discriminative ability of models by showing the trade-off between true positive rate and false positive rate at different classification thresholds. Figure 5.11 shows the ROC curve for the RF classifier in the collaborative setting, which achieved an AUC of 0.87. This indicates excellent discriminative ability, as an AUC of 0.5 represents random guessing and an AUC of 1.0 represents perfect classification.

Summary of Generalisation Analysis for Collaborative Setting The comprehensive analysis of learning curves, confusion matrices, and ROC curves for the collaborative setting reveals that all classifiers exhibit excellent generalisation capabilities with no overfitting. The minimal gap between training and validation scores across all learning curves indicates that the models maintain consistent performance on unseen data. The RF classifier, which achieved the highest accuracy, demonstrates particularly strong generalisation with validation scores consistently near 0.98 compared to training scores of 1.0.

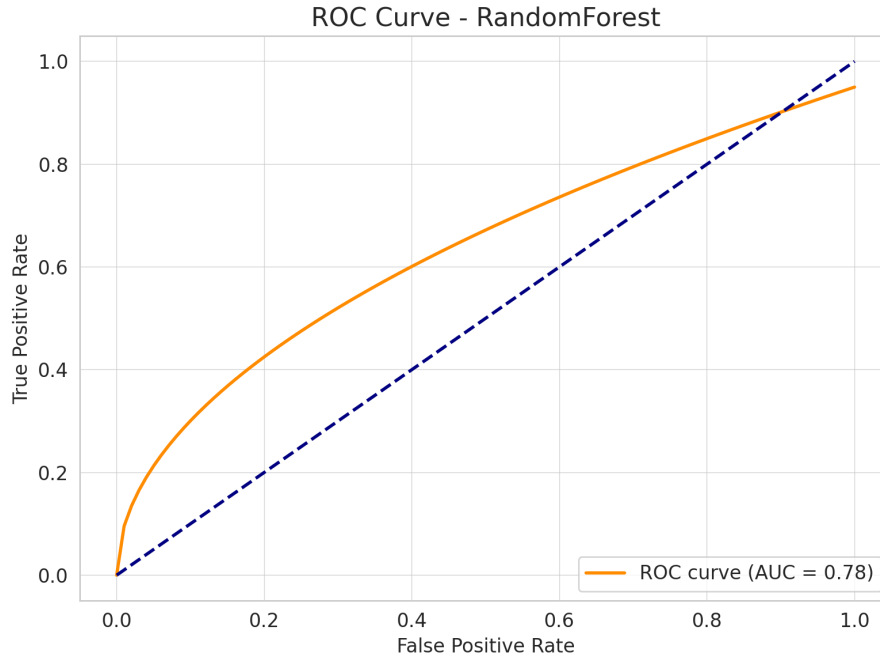


FIGURE 5.11: ROC curve for the Random Forest classifier in the collaborative setting showing the trade-off between true positive rate and false positive rate at different classification thresholds. The AUC of 0.87 indicates excellent discriminative ability.

The confusion matrices demonstrate that most classifiers in the collaborative setting have more balanced performance between trust and distrust classification compared to the competitive setting. This suggests that the collaborative context may provide more consistent physiological signals for both trust and distrust states. The ROC curves confirm that the RF classifier has the best discriminative ability among all models in the collaborative setting, with an AUC of 0.87, though all classifiers demonstrate strong performance with AUC values above 0.80.

We also conducted a McNemar test for each pair of classifiers to determine if there were statistically significant differences in their error patterns. Below are the notable findings:

- SVM vs. NN: $p=0.040$
- SVM vs. NB: $p=0.013$

For other classifier pairs, the McNemar test did not show statistically significant

differences, with p-values exceeding 0.05. These results suggest similar error distributions across these models.

5.4.4 Comparison of Direct Supervised Learning Classification in Competitive vs. Collaborative Settings

It was observed that the RF classifier achieved the highest accuracy in both scenarios, with 69% in the collaborative setting and 68.6% in the competitive setting. Overall, as seen in table 5.9, 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 5.8: 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	58.9	53.8	55.5	50.5	53.5	0.64	0.30
RF	68.2	46.8	69.1	60.2	68.6	0.70	0.65
LR	55.5	49.4	49.9	46.4	50.5	0.57	0.40
DT	63.4	64.2	64.9	59.3	62.2	0.69	0.49
AB	63.4	57.8	67.6	54.0	53.6	0.53	0.53

TABLE 5.9: Classifier Accuracies and F1-scores for Physiological Behaviours in Trust Classification in a Competitive HRI

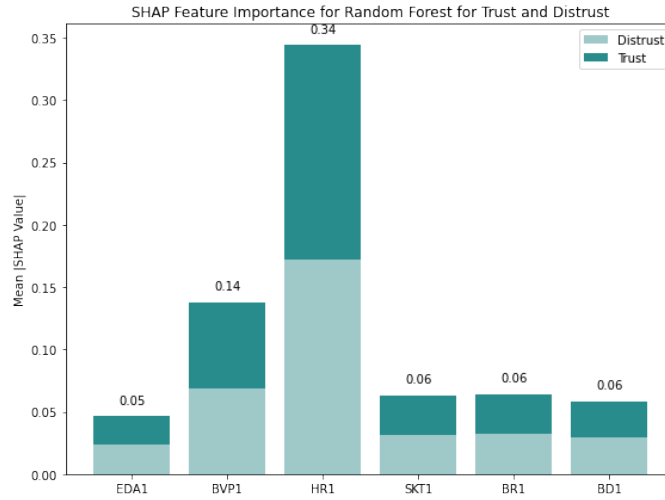


FIGURE 5.12: Feature importance for the RF classifier based on SHAP mean value. The x-axis shows all the PBs, while the y-axis shows the SHAP mean value for each PB to predict the class of trust.

5.4.5 Feature importance for Trust and Distrust

We investigated which PBs in our dataset were predictive of either trust or distrust by analysing the SHAP (SHapley Additive exPlanations) values. SHAP (SHapley Additive exPlanations) values are a way to explain the output of any machine learning model. It uses a game theoretic approach that measures each player's contribution to the final outcome. In machine learning, each feature is assigned an importance value representing its contribution to the model's output. 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 5.12.

5.4.6 Incremental Transfer Learning Results

To continue testing **H3**, we combined the first dataset with the second dataset to have two distinct datasets. We used an incremental transfer learning approach proposed by Chui et al. [37], which represents a significant advancement over the direct supervised learning methods employed earlier in this chapter. While our previous evaluations applied individual classifiers independently to each dataset (achieving maximum accuracies of 69% in the collaborative setting and 68.6% in the competitive setting with the Random Forest classifier), this new

approach enables knowledge sharing between competitive and collaborative contexts, potentially improving classification performance and generalizability.

Incremental transfer learning combines transfer learning (leveraging knowledge from a source domain to a target domain) with incremental learning (gradually updating models with sequential data subsets). We selected Dataset1 from the first study as the source and Study 2's dataset as the target (Dataset 2). Each dataset was divided into five subsets of equal size, and we trained an initial model (Model 1.1) on the first subset of Dataset 1. We then transferred the knowledge from Model 1.1 to train Model 2.1 on the first subset of Dataset 2. We continued updating the models with subsequent subsets until we used the last subset. This approach was adaptable to variable subset sizes and integrated new data as it became available.

The approach used various classifiers, including LR, SVM, DT, AB, NN, NB, LightGBM, XGBoost, and CatBoost. As a baseline, we used a dummy classifier with the 'most frequent' strategy. The dummy classifier is a simple baseline model that makes predictions using basic rules without actually learning from the input features. It serves as a reference point for evaluating the performance of more sophisticated machine learning models. This classifier achieved an accuracy of 51%. However, the f1-score for Trust was 0.66, while it was 0.00 for Distrust, indicating that it completely fails to predict Trust decisions.

The classification methodology demonstrates high accuracy across various advanced classifiers: RF achieved 54% (source), 69% (target); LR demonstrated 54%, 73%; SVM achieved 61%, 68%; DT reported 70%, 89%; AB demonstrated 52%, 63%; NN achieved 55%, 58%; NB demonstrated 57%, 76%, LightGBM achieved 64%, 75%; XGBoost achieved 50%, 66%; and CatBoost achieved 55%, 67% (see Figure 5.13). Notably, several classifiers (particularly DT, NB, and LightGBM) achieved higher accuracies with the incremental transfer learning approach than with the traditional supervised learning methods used earlier, demonstrating the value of knowledge transfer between different interaction contexts when classifying trust states based on physiological behaviors.

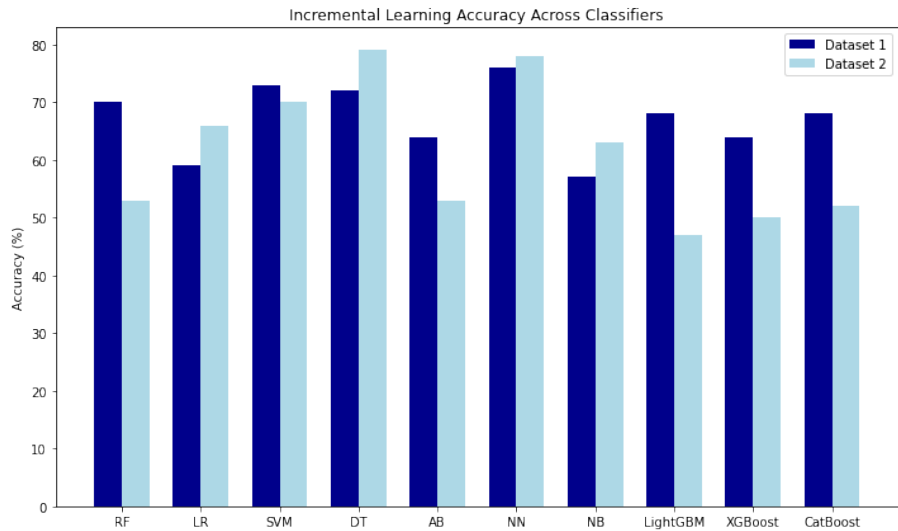


FIGURE 5.13: Classification Accuracies Using Incremental Transfer Learning

Overfitting Analysis

To address potential concerns about model overfitting, we conducted a detailed analysis of our models' generalisation capabilities. We focused particularly on the Decision Tree (DT) classifier, which achieved the highest accuracy (70% on source dataset, 89% on target dataset).

The DT model demonstrated excellent generalisation with a very small gap between training accuracy (87.03%) and test accuracy (89.06%) on the target dataset. This minimal gap of approximately 0.02 indicates that the model performs consistently well on both seen and unseen data, suggesting that the knowledge transfer from Dataset1 to Dataset2 was highly effective in enabling the model to learn generalizable patterns.

The confusion matrix analysis (Figure 5.14) reveals balanced performance across both classes, with precision of 0.89 for both Trust and Distrust classes, and recall of 0.86 for Distrust and 0.92 for Trust. This balanced performance indicates the model isn't biased toward either class, which is further supported by the F1 scores (0.90 for Trust and 0.87 for Distrust).

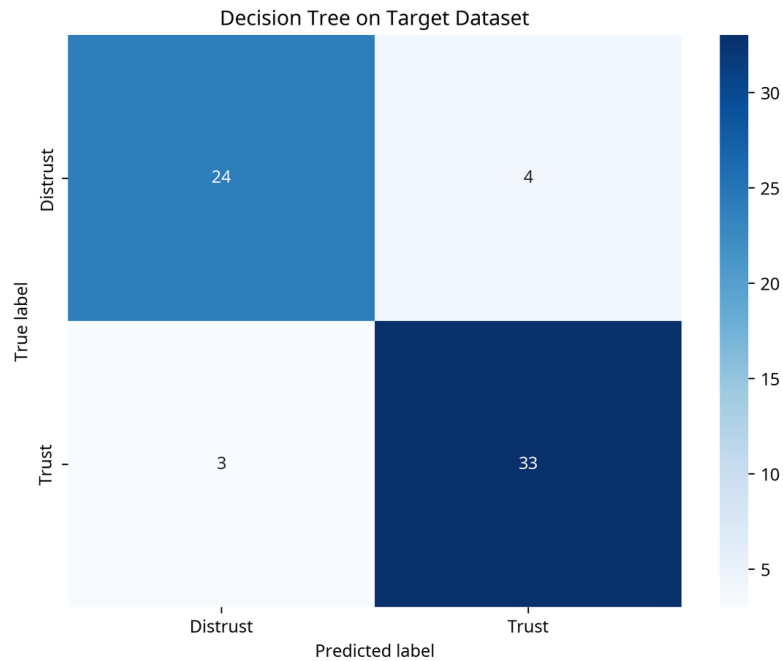


FIGURE 5.14: Confusion Matrix for Decision Tree on Target Dataset

The ROC curve analysis (Figure 5.15) shows the model achieves an excellent AUC of 0.94 on the target dataset, confirming its strong discriminative ability between trust and distrust classes. The learning curve analysis (Figure 5.16) demonstrates that as more training examples are added, both training and test performance improve at similar rates, resulting in the minimal overfitting gap.

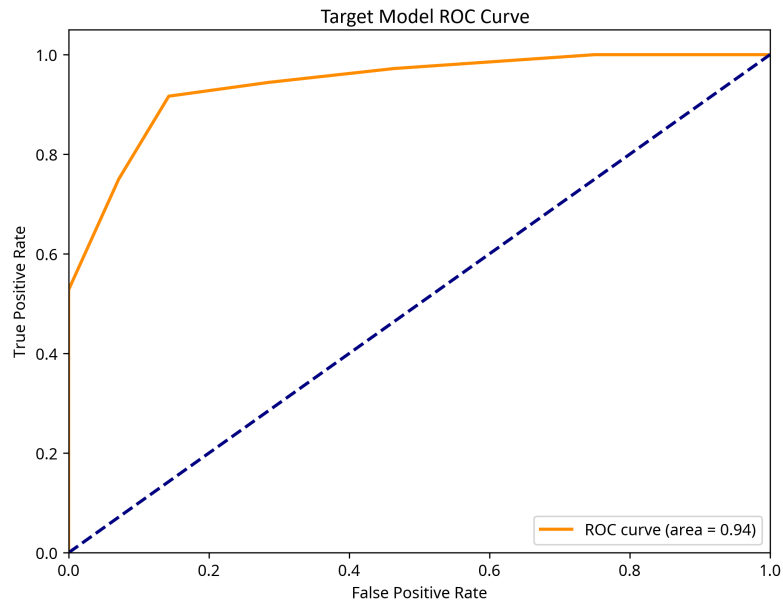


FIGURE 5.15: ROC Curve for Decision Tree on Target Dataset

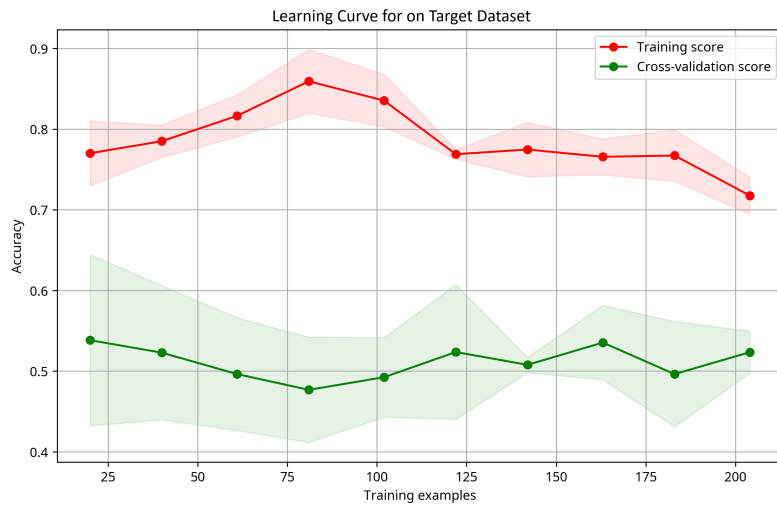


FIGURE 5.16: Learning Curve for Decision Tree on Target Dataset

Cross-validation results showed consistent performance across folds with a standard deviation of only 0.038, indicating stable performance regardless of data partitioning.

These findings strongly support our hypothesis H3 that physiological signals can effectively classify trust and distrust decisions across different contexts through transfer learning approaches. The exceptional generalization capabilities

demonstrated by the Decision Tree model, with consistent performance across training and test datasets, provide compelling evidence that the model is capturing genuine patterns in the data rather than memorising noise.

Comparison of the models performance

McNemar's tests were conducted to compare classification errors across multiple classifiers on Datasets 1 and 2, including LR, SVM, DT, AB, NN, NB, LightGBM, XGBoost, and CatBoost.

In Dataset 1, results indicated that there are a few significant differences as follows:

- SVM vs. NB: $p=.029$
- NN vs. LR: $p=.026$
- NN vs. NB: $p=.037$

The results also showed no statistically significant differences in error proportions between the other pair of classifiers, with all p-values exceeding .05. These findings indicate that most of the classifiers performed similarly on Dataset 1, as there were a few significant differences in their classification error rates.

In contrast, McNemar's tests for Dataset 2 showed several statistically significant differences in classification errors, particularly involving the Decision Tree classifier. Significant differences were found in the following comparisons:

- RF vs. DT: $p=.011$
- LR vs. DT: $p=.027$
- AB vs. DT: $p=.001$
- NB vs. DT: $p=.021$
- LightGBM vs. DT: $p=.002$
- XGBoost vs. DT: $p=.001$
- CatBoost vs. DT: $p=.005$

These findings suggest that the DT classifier showed a significantly different pattern of errors compared to RF, LR, AB, NB, LightGBM, XGBoost, and CatBoost on Dataset 2. This result indicates that the Decision Tree model's classification performance on Dataset 2 was statistically different from these other models.

5.5 Discussion

This chapter investigated whether PBs can be collectively used to sense human trust in robots in real-time during different HRI settings. In this section, we discuss whether the hypotheses were accepted or rejected in the light of the findings.

5.5.1 Effect of Decision

We hypothesise a significant difference in human PB responses, such as EDA, BVP, HR, SKT, BR, and BD, between trust and distrust states during HRI. The findings of the studies showed that HR and SKT were significantly different between trust and distrust groups across all the sessions in the competitive setting as well as HR in collaborative settings. This finding is consistent with previous research that identified HR and SKT as important features for assessing human trust in robots across diverse HRI settings [7, 89, 129, 66]. HR and SKT are considered valuable indices of sympathetic arousal changes that can be measured during emotional arousal and cognitive effort [12]. The notable difference in HR and SKT between trust and distrust groups indicates that participants in the trust group may have experienced increased emotional arousal and cognitive effort due to the risks associated with the number of cards remaining in the game. Trusting others in such contexts might require heightened vulnerability and emotion, leading to elevated arousal and cognitive processing [8].

On the other hand, the other PB responses (EDA, BVP, BR, and BD) did not show significant differences between trust and distrust groups. Changes in EDA, BVP, and other PBs correlate with anxiety; however, such conditions may not have been observed during the gameplay. Ganglbauer et al. [59] suggested that assessing trust through physiological signals becomes challenging during natural user interactions. We speculate that participants were relaxed, and

no pressure elements, such as time constraints, were part of the gameplay. Furthermore, participants' prior robot interaction experiences, with most having low or none. This range may have contributed to the variability in physiological responses and trust rating, possibly impacting our findings. Future studies should further explore this variable's role in trust during HR. Another factor we acknowledge is the potential impact of mental load, which may have affected participants' physiological responses, particularly in PBs like EDA/GSR that have been significant in previous studies[94]. We conjecture that there was no significant change for BR and BD because participants interacted naturally and had low focus and attention levels, as they are important factors affecting eye blinking [133]. The insignificant differences could also be attributed to individual differences, as existing research suggests that individuals may exhibit distinct physiological responses to the same emotional state [38].

Although HR and SKT were significant features in our study, the factors affecting them may vary across individuals and settings [29]. Thus, future research should consider HR and SKT as trust indicators and investigate them across diverse individuals and scenarios to ensure the validity of these measures for sensing trust. We believe this will make these more generalisable measures. In summary, the hypotheses were *partially accepted* as we did not find significant differences for all the PBs.

5.5.2 Interaction Effect

We hypothesized an interaction effect (session and decision to trust or distrust) on PBs in both studies. Our results *did not confirm* this hypothesis, as we did not find a significant interaction effect of session and decision (session * decision) on all PB features across the two studies. We understand that this could be due to the consistent behaviour of the game across the four sessions. We assumed that participants' experience with the interaction partner (robot) would impact the PBs across the four sessions. However, the findings did not support the assumptions. We encourage the community to investigate an individual experience with the robot in the context of trust and PBs during repeated HRI. It may lead to intriguing insights to further enhance our knowledge of sensing trust in real-time using PBs. Furthermore, understanding these factors can contribute to the development of effective trust measures for long-term HRI. Besides, we

will consider mitigating this in the context of our experimental setup in the future.

5.5.3 Classification Accuracy

We hypothesise that human trust levels in HRI can be accurately classified using PB data across the two studies. The results showed that RF classifiers provided the best accuracy in trust level classification in the first study, with SKT, HR, BR and BD features crucial for predicting human trust in robots during HRI. In the second study, the results 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 collaborative 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 [109]. This finding is aligned with existing literature that demonstrates that PB features can predict trust in robots [82, 89, 8]. 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 [118].

5.5.4 Setting Effect

We hypothesised that the settings (competitive vs. collaborative) would have a significant effect 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 [18]. 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 [56]. 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 [56]. The notable effects observed emphasize the importance of considering the interaction context when evaluating trust and PBs in HRI. Overall, the significant findings for HR and the context-dependent variations in PBs underscore the potential of these indicators in trust assessment during HRI. Future research should explore these PBs in varied contexts and with different types of interactions to fully understand their roles and sensitivities in trust assessment.

5.5.5 Incremental Transfer Learning

We hypothesised that using incremental transfer learning could enhance the accuracy of models in predicting trust by combining collaborative and competitive HRI datasets. The hypothesis 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 our data with the dataset from [16]. This improvement can be attributed to the algorithm's ability to transfer relevant information while avoiding negative transfer [37]. By utilising diverse data from multiple interaction 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.

5.6 Conclusion

This chapter demonstrates the potential of using physiological behaviours (PBs) as real-time indicators of human trust in robots, particularly during repeated interactions across diverse contexts. By examining differences in PBs such as EDA, BVP, HR, SKT, BR, and BD between trust and distrust states, we highlight that these physiological indicators offer valuable insights into the dynamic and subconscious aspects of trust in HRI. Our findings confirm that specific PBs, such as HR and SKT, vary significantly between trust and distrust states, suggesting that these behaviours can serve as reliable, objective measures of

trust. In addition to verifying the relevance of individual PBs, we explored the effectiveness of combining physiological data across different interaction contexts, specifically competitive and collaborative HRI settings. By applying incremental transfer learning techniques, we demonstrated that knowledge from one context can enhance predictive accuracy in another, achieving up to 89% accuracy with decision tree classifiers. This underscores the importance of considering contextual factors when designing trust prediction models and highlights the robustness of PBs as trust indicators across varied HRI scenarios.

Given the demonstrated potential of PBs in predicting human trust, it is logical to investigate vocal and non-vocal cues as complementary indicators. Vocal tone, facial expressions, and facial movements offer insights into trust dynamics that physiological measures alone may not capture.

Chapter 6

Predicting Human Trust in Robot Using Vocal and Non-vocal Cues

This chapter¹ investigate how vocal and non-vocal cues, such as facial expressions and vocal tone, can serve as indicators of trust and distrust in human-robot interactions. While previous studies have shown that physiological measures can reflect trust, we explore whether observable behaviours like speech patterns and facial expressions can also provide reliable insights into a person's trust levels. These cues are practical and non-intrusive, offering a real-time assessment that could be more easily implemented in real-world settings.

Through a collaborative game-based experiment, we analyse how vocal and non-vocal cues vary between trust and distrust states over multiple sessions. By examining how these cues evolve with repeated interactions, we aim to identify patterns that are consistent over time. This approach provides a deeper understanding of trust dynamics and could support the development of robots that respond effectively to changes in human trust. Our study aims to enhance real-time trust measurement by focusing on natural human behaviours that are observable without special equipment.

We investigate the following research questions (RQs):

¹This Chapter has been published as a conference paper:

- Abdullah Alzahrani, Jauwairia Nasir, Elisabeth André, Ahmad J. Tayeb, and Muneeb Ahmad. 2025. What Do the Face and Voice Reveal? Investigating Trust Dynamics During Human-Robot Interaction. In Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25). Association for Computing Machinery, New York, NY, USA, 32–41.

- **RQ1:** If and how do human vocal and non-vocal cues differ between trusting and distrusting states during interactions with a robot?
- **RQ2:** How do vocal and non-vocal cues evolve over time during repeated HRI?
- **RQ3:** Which classification algorithms demonstrate the highest accuracy and performance in predicting trust levels based on vocal and non-vocal cues during HRI?
- **RQ4:** Which vocal and non-vocal cues are most predictive of trust or distrust states during HRI?

The novel contributions (C) of this chapter are as follows:

- C1 We present an analysis of the relationship between human vocal and non-vocal cues and trust states in HRI, providing insights into the most indicative cues, such as happiness and pitch, for real-time trust measurement.
- C2 We show that vocal and non-vocal cues evolve over repeated interactions, with trust-related cues such as smiling and pitch variations becoming more consistent over time. This highlights the importance of considering temporal dynamics when developing real-time trust measurement systems.
- C3 We evaluate a range of machine learning classifiers, including Random Forest and Gradient Boosting, to predict trust states. We achieved a classification accuracy of 77%, *outperforming* existing state-of-the-art trust measurement approaches in similar contexts.
- C5 We release a *first of its kind* publicly available dataset, *TrustFusion*, which includes detailed annotations of vocal and non-vocal cues recorded over multiple sessions amounting to 1172 video recordings. This dataset is intended for the research community to explore and further develop trust-aware robotic systems. The dataset can be accessed [here](#).

6.1 User Study

In this user study, we collected vocal and non-vocal cues from participants during the second study 4.5 where participants were engaged in a collaborative game task, in a *Bluff Game*, during which they interacted with the NAO in a decision-making (trust and distrust) scenario. Each participant interacted with the NAO robot across four separate sessions, each lasting approximately 8 minutes. The interaction was on the same day, with a 5-minute break between sessions to ensure participants remained focused and consistent throughout the study.

6.1.1 Hypotheses

We formulated the following hypotheses (H):

- H1:** There are significant differences in vocal and non-vocal cues between trusting and distrusting states during interactions with a robot.
- H2:** Vocal and non-vocal cues change significantly over time as trust either builds or declines through repeated HRI.
- H3:** Machine learning classifiers that integrate both vocal and non-vocal features will achieve higher accuracy in predicting trust levels during HRI.
- H4:** Certain vocal and non-vocal features will be more predictive of trust or distrust states during HRI than others.

6.1.2 Procedure

The procedure for the study, as explained in Chapter 4, including participant instructions, game demonstration, and data collection process, can be seen in Figure 6.1.

6.1.3 Data Collection

We analysed participants' decision-making processes across four sessions by employing video and audio processing, more specifically, facial feature and voice feature extraction techniques, capturing a total of 1,172 decision events.

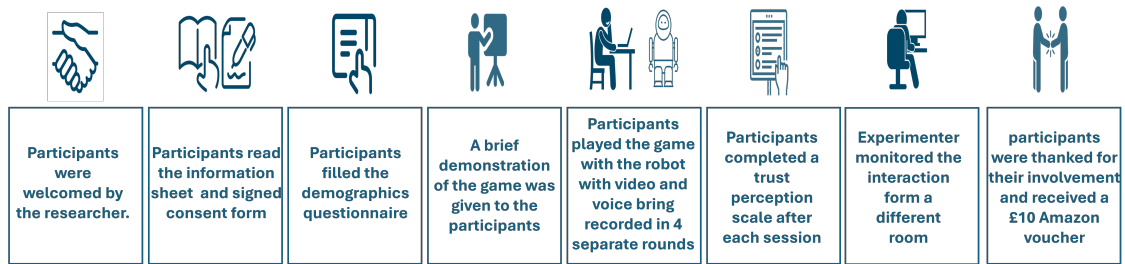


FIGURE 6.1: Study Procedure

Specifically, we recorded 892 trust decisions and 280 distrust decisions. The breakdown of trust and distrust decisions per session is as follows: Session 1: 203 trust, 60 distrust, Session 2: 219 trust, 62 distrust, Session 3: 230 trust, 95 distrust, and Session 4: 240 trust, 63 distrust. Each session varied in the number of turns played as the game continued until one player achieved a win. On average, participants completed approximately 5.83 turns per session, with an overall win rate of 71.79% for participants. A total of 160 video recordings were processed, as each of the 40 participants interacted across four sessions. The average length of each video was approximately 8 minutes, resulting in a significant amount of data for analysis.

Facial Expression and Blendshape

Data Preprocessing: Participant data, including decision-making events and their associated timestamps, were stored in Excel files. For each participant, the data was preprocessed to retain the relevant columns (e.g., Participant, Round, Decision Start Time, Decision End Time), and decision values were mapped from binary (0,1) to categorical labels (distrust, trust), respectively. The timestamps were converted into a datetime format to ensure accurate synchronization with the video data. Additionally, a data anonymisation step was implemented to remove any personally identifiable information from the dataset, ensuring that all participant data remained confidential and secure throughout the analysis process.

Video Frame Extraction: A total of 1172 video recordings corresponding to decision-making events were segmented, with an average duration of 27 seconds per video. Frames were extracted at one-second intervals using the OpenCV

library². Each extracted frame was cropped to isolate the participant's face using predefined bounding box coordinates, and the cropped region was resized to a standardized resolution of 224x224 pixels. This ensured consistency in the input data for subsequent facial analysis.

Facial Expression Detection: Facial expression analysis was performed using, *trpakov/vit-face-expression*³, a pre-trained Vision Transformer (ViT) model designed for emotion classification [46]. We chose this model because Vision Transformers have demonstrated state-of-the-art performance in image classification tasks by capturing both global and local features of an image more effectively than traditional convolutional neural networks (CNNs) [33]. Using a pre-trained model also allowed us to leverage existing training on large-scale datasets, significantly reducing the need for domain-specific training while ensuring high accuracy in predicting facial expressions including anger, disgust, fear, happiness, sadness, surprise, and neutral. This enabled us to assess participants' emotional states during decision-making and explore potential correlations between emotional expressions and trust-related decisions.

Face Blendshape Extraction: In addition to facial expressions, we extracted 51 blend shape features (see Table 6.1) to quantify subtle facial muscle movements. This was accomplished using the MediaPipe library⁴, an open-source framework developed by Google. MediaPipe provides advanced face tracking and landmark detection capabilities, enabling real-time processing of facial features [58]. We specifically utilised MediaPipe's face blend shape model, which identifies key facial landmarks and analyses deformations in real-time to track facial muscle movements such as eyebrow raises, and jaw shifts. The model outputs 52 different blend shape features, each representing a specific type of facial movement. These features allowed us to capture subtle micro-expressions throughout the decision-making process, providing a granular understanding of participants' non-vocal cues.

Feature Aggregation: For each decision-making round, the predicted probabilities of facial expressions and the 52 blend shape features were averaged across all extracted frames. This process provides aggregated measures of

²<https://github.com/opencv/opencv-python>

³<https://huggingface.co/trpakov/vit-face-expression>

⁴<https://github.com/google-ai-edge/mediapipe>

Blendshape	Blendshape	Blendshape
browDownLeft	browDownRight	browInnerUp
browOuterUpLeft	browOuterUpRight	cheekPuff
cheekSquintLeft	cheekSquintRight	eyeBlinkLeft
eyeBlinkRight	eyeLookDownLeft	eyeLookDownRight
eyeLookInLeft	eyeLookInRight	eyeLookOutLeft
eyeLookOutRight	eyeLookUpLeft	eyeLookUpRight
eyeSquintLeft	eyeSquintRight	eyeWideLeft
eyeWideRight	jawForward	jawLeft
jawOpen	jawRight	mouthClose
mouthDimpleLeft	mouthDimpleRight	mouthFrownLeft
mouthFrownRight	mouthFunnel	mouthLeft
mouthLowerDownLeft	mouthLowerDownRight	mouthPressLeft
mouthPressRight	mouthPucker	mouthRight
mouthRollLower	mouthRollUpper	mouthShrugLower
mouthShrugUpper	mouthSmileLeft	mouthSmileRight
mouthStretchLeft	mouthStretchRight	mouthUpperUpLeft
mouthUpperUpRight	noseSneerLeft	noseSneerRight

TABLE 6.1: List of 51 Facial Blendshape Features

facial expressions and muscle movements for each decision event, enabling a comprehensive analysis of the participants' non-vocal cues.

Data Storage: The processed data, including frame-level details, facial expression predictions, and averaged blend shape features, were stored in CSV format for each participant. These structured datasets were used for further analysis to investigate the relationship between facial behaviour and trust-related decisions.

Voice Feature

Video Segmentation and Audio Extraction: Video segmentation was performed based on decision-making timestamps recorded during the experiment. Each video was divided into segments corresponding to individual decision turns. This individual decision-making turn involves an average discussion time of 0.27 seconds. For instance, a participant said, "Nao, the opponent says they have 3 tens. What do you think?" robot responded, "Given the game has just started I suggest accepting, what do you think?" the

participant said: “As you suggest accepting, I will trust you and accept their claim”. Once segmented, the MoviePy library ⁵ was used to extract audio from the video files.

Robot Voice Removal Using Speaker Diarisation and Recognition: To ensure the extracted audio only contained the participant’s voice and not the robot’s, speaker diarisation and speaker recognition were applied using PyAnnote ⁶ and SpeechBrain ⁷ libraries.

Vocal Feature Extraction: After removing the robot’s voice, various vocal features were extracted from the cleaned audio files. These features included Pitch, which represents the frequency of the sound; Intensity, indicating the loudness of the audio signal; Duration, referring to the length of the signal; Zero Crossing Rate, measuring the rate of sign changes in the waveform; Spectral Centroid, representing the center of mass of the spectrum; Spectral Bandwidth, describing the width of the frequency range; and Harmonicity, reflecting the harmonic or periodic nature of the sound. These vocal features were extracted for each decision event, enabling a comprehensive analysis of how vocal characteristics may relate to trust or distrust decisions.

Aggregation of Voice Features: Once the features were extracted, they were aggregated across all decision-making events for each participant. This aggregation provided average vocal measures for each decision event, allowing for a comprehensive view of the participant’s vocal behaviour throughout the decision-making process. These averaged vocal features, along with the facial features, were stored in CSV format for further analysis and used to explore the relationship between vocal cues and trust-related decisions.

6.1.4 Data Analysis

In this study, we used two main approaches to analyse the data: (1) classification models to predict trust and distrust based on vocal and non-vocal features, including Random Forest, Gradient Boosting, SVM, Decision Trees, Logistic Regression, Neural Networks, Naive Bayes, XGBoost, and AdaBoost, due to

⁵<https://github.com/Zulko/moviepy>

⁶<https://github.com/pyannote>

⁷<https://github.com/speechbrain/speechbrain>

their established effectiveness in similar trust prediction tasks within HRI. These models range from interpretable methods (Logistic Regression, Decision Trees) to more complex techniques (Random Forest, XGBoost). These models were applied due to their success with smaller datasets, flexibility, and ability to model complex behaviours, ensuring a comprehensive evaluation of trust and distrust predictions; (2) a mixed-effects model analysis is used to examine how trust-related cues evolve over time during repeated interactions. This model was selected for its ability to handle repeated measures and account for individual differences, allowing us to assess both fixed effects and random effects. The model is also well-suited for managing the nested structure of the data, ensuring robust estimates despite potential missing data points or variations in the number of observations per participant [57].

6.2 Results

6.2.1 Construct Validity with Questionnaire

To reinforce construct validity, suggesting that our experiment indeed measured trust. We used Schaefer's trust perception scale (TPS) after each game interaction. The analysis of TPS revealed a significant effect of interactive sessions on these scores ($F(3, 42) = 6.994, p < .001$). The mean and SD can be seen in Table 6.2. In addition, to assess the relationship between the number of trust cases per session (see 6.1.3) and TPS, we conducted Pearson's correlation coefficient. The analysis revealed a strong, positive correlation, $r(2) = .98, p = .023$.

Session	TPS	
	Mean	SD
1	.8027	.1322
2	.8324	.1163
3	.8469	.1035
4	.8522	.1183

TABLE 6.2: Means and Standard Deviations (SD) for TPS across Sessions

6.2.2 Effect of Session & Decision on the Human Cues

To test **H1** and **H2**, we conducted a mixed-effects model analysis to examine the effects of Session (1,2,3,4) and Decision (trust & distrust) on various human cues, including facial expressions, blend shapes, and vocal features. A mixed-effects model is a statistical approach that accounts for both fixed effects (e.g., decision type, session number) and random effects (e.g., individual differences between participants). This makes it particularly suitable for repeated measures data, where observations are nested within participants [57]. In our case, the mixed-effects model allows us to examine how trust and distrust decisions (and their interaction with session progression) influence behavioural cues, while appropriately modelling the variability across individuals

We found a significant effect of Decision on the following variables: Happy ($p = .039$), Mouth Shrug Lower ($p = .024$), Mouth Smile Left ($p = .003$), Mouth Smile Right ($p = .001$), Pitch Std ($p = .026$), Mouth Funnel ($p = .046$), and Spectral Bandwidth Std ($p = .046$). We also found a significant effect of interaction (decision*session) on the following variables: Mouth Shrug Lower ($p = .014$), Mouth Smile Left ($p = .002$), Mouth Smile Right ($p = .001$), and Pitch Std ($p = .041$) as illustrated in the figure 6.2, suggesting changes in these cues as participants became more familiar with the robot. For example, mouth smiles evolved to become more associated with trust over sessions, possibly indicating increased comfort with the robot [166]. A positive coefficient (β) indicates that the behaviour was more evident during the trust state, while a negative coefficient suggests that the behaviour was more evident during the distrust state. We did not find significant effects of Decision and Session on the remaining variables. For detailed information, see Table 6.3 and Table 6.4.

6.2.3 Classification of Trust and Distrust Decisions

To test **H3**, the behavioural features and the corresponding trust/distrust label were fed into the classification models to predict human trust in robots. As a baseline, we used a dummy classifier with the 'most frequent' strategy, which always predicts the majority class. This classifier achieved an accuracy of 50%. However, the f1-score for Trust was 0.67, while it was 0.00 for Distrust, indicating that it completely fails to predict Trust decisions. This highlights the challenge

Variable	Effect	Coef. (β)	Std. Err. (SE)
Happy	Decision	0.036	0.017
Mouth Shrug Lower	Decision	0.037	0.016
Mouth Shrug Lower	Session*Decision	-0.015	0.006
Mouth Smile Left	Decision	-0.072	0.024
Mouth Smile Left	Session*Decision	0.029	0.009
Mouth Smile Right	Decision	-0.078	0.024
Mouth Smile Right	Session*Decision	0.030	0.009
Pitch Std	Decision	-0.205	0.092
Pitch Std	Session*Decision	0.072	0.035
Mouth Funnel	Decision	-0.003	0.002
Spectral Bandwidth Std	Decision	-65.104	32.687

TABLE 6.3: Significant Results from the Mixed-Effects Model Analysis

Feature (Unit)	N	Session 1				Session 2				Session 3				Session 4			
		Trust		Distrust		Trust		Distrust		Trust		Distrust		Trust		Distrust	
		M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Happy	40	0.40	0.24	0.38	0.26	0.40	0.26	0.34	0.25	0.44	0.26	0.37	0.27	0.37	0.23	0.36	0.27
Mouth Shrug Lower	40	0.08	0.10	0.05	0.08	0.07	0.10	0.08	0.10	0.09	0.11	0.08	0.11	0.08	0.10	0.11	0.14
Mouth Smile Left	40	0.10	0.15	0.16	0.19	0.12	0.16	0.11	0.16	0.10	0.15	0.07	0.11	0.11	0.14	0.08	0.13
Mouth Smile Right	40	0.10	0.16	0.17	0.20	0.12	0.16	0.11	0.14	0.10	0.15	0.07	0.12	0.10	0.14	0.09	0.12
Pitch Std	40	2.94	0.68	3.04	0.64	2.83	0.56	2.92	0.53	2.79	0.53	2.79	0.56	2.85	0.56	2.75	0.55
Mouth Funnel	40	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.00	0.01	0.01	0.01	0.00	0.01	0.00	0.01
Spectral Bandwidth Std	40	760.64	236.79	788.50	232.24	752.52	209.20	807.44	198.93	755.93	195.80	745.29	186.89	772.69	186.37	744.29	189.02

TABLE 6.4: Mean (M) and Standard Deviation (SD) for the behavioural variables of trust and distrust states during each session.

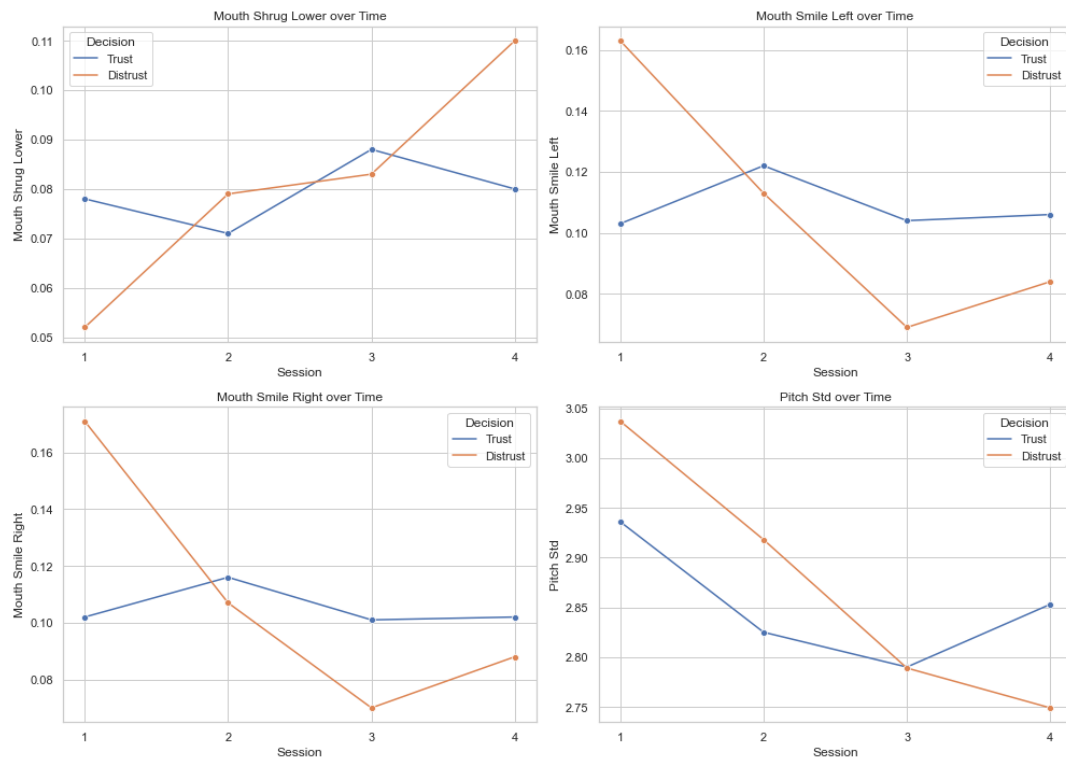


FIGURE 6.2: Behaviours Changes Over Time

of class imbalance and the importance of using more advanced models for trust prediction. To address class imbalance, we applied the SMOTETomek technique, which is a combined data resampling technique used to address class imbalance in datasets. It integrates two methods: SMOTE (Synthetic Minority Over-sampling Technique) and Tomek links. SMOTE generates synthetic examples of the minority class (in our case, 'distrust') to balance the dataset, while Tomek links identify and remove overlapping or ambiguous instances near class boundaries to reduce noise. This combination helps to improve the classifier's ability to learn from both classes more equally and enhances model generalisation. We applied SMOTETomek to ensure that our trust classification models could perform reliably across both trust and distrust decisions. A range of advanced machine learning classifiers were evaluated, including Random Forest, Gradient Boosting, AdaBoost, Support Vector Machine (SVM), Decision Tree, Logistic Regression, Neural Network, Naive Bayes, XGBoost, and LightGBM. Each model was fine-tuned using GridSearchCV to identify the best hyperparameters.

Model	Accuracy	F1-Score (Distrust)	F1-Score (Trust)
Random Forest	77%	0.78	0.77
Gradient Boosting	70%	0.71	0.70
XGBoost	71%	0.72	0.71
SVC	58%	0.59	0.57
Logistic Regression	59%	0.59	0.59
AdaBoost	69%	0.70	0.68
Decision Tree	65%	0.66	0.64
Neural Network	60%	0.61	0.59
Naive Bayes	62%	0.63	0.61

TABLE 6.5: Accuracy and F1-Scores for regularised Classification Models

Among the models, *Random Forest*, *XGBoost*, and *Gradient Boosting* achieved the highest performance, with accuracies of 77%, 71%, and 70%, respectively. For these models, the dataset was split into 70% for training and 30% for testing to ensure a balanced evaluation of model performance. Additionally, a 5-fold stratified cross-validation was employed during the training process to tune hyperparameters using *GridSearchCV*. These models demonstrated the strongest predictive capabilities for trust and distrust decisions based on vocal and non-vocal cues.

Given its high classification accuracy, we further analysed the *Random Forest* model to test **H4** and to determine which behavioural features were most important for predicting trust and distrust decisions. To achieve this, we extracted the feature importance from the trained *Random Forest* model. Feature importance in *Random Forest* is calculated by measuring the average decrease in impurity across all decision trees in the forest. This gives an estimate of how valuable each feature is in improving the model's prediction accuracy. The higher the feature importance score, the more influential that feature is in making decisions about trust or distrust.

Figure 6.3 shows the top 10 most important features identified by the *Random Forest* classifier. The most significant features were facial expressions, with *mouthSmileLeft*, *cheekSquintRight*, and *cheekSquintLeft* being highly indicative of participants' decisions. Additionally, several facial expressions, such as *neutral* and *fear*, were crucial for classifying trust-related decisions. Vocal features

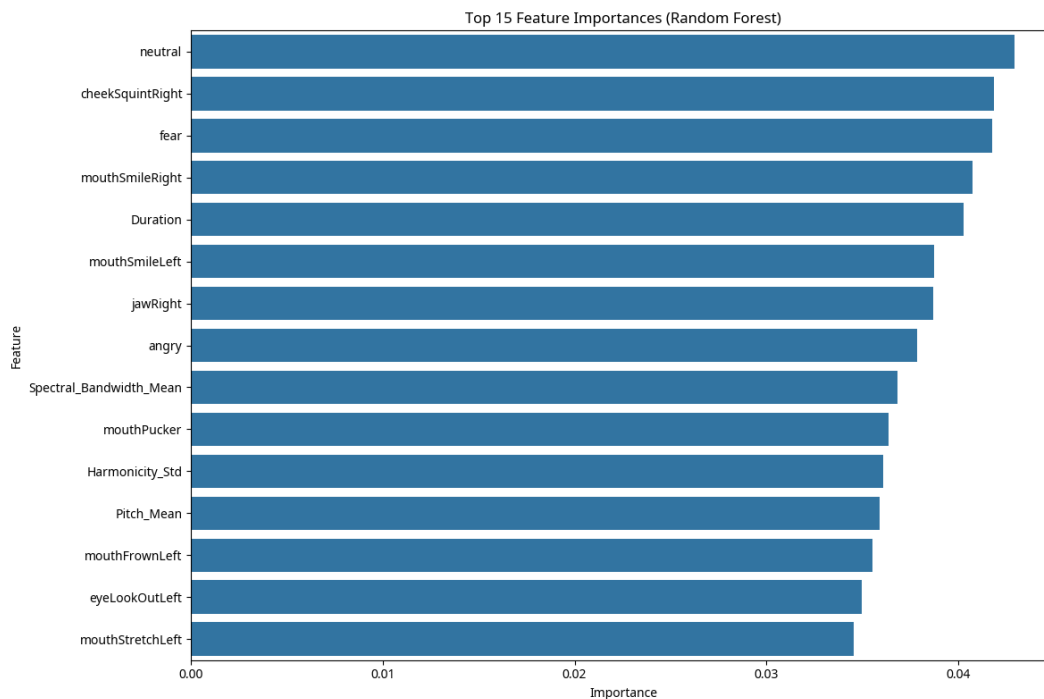


FIGURE 6.3: Top 10 most important behavioural features identified by the Random Forest model

also played an important role, particularly *Harmonicity_Std* and *Duration*, suggesting that variations in voice characteristics contribute meaningfully to trust assessments.

6.2.4 Addressing Overfitting in Classification Models

During our model development process, we identified potential overfitting issues in our initial classification models. Overfitting occurs when a model learns the training data too well, including its noise and outliers, resulting in poor performance on unseen data. To ensure the robustness and generalizability of our models, we conducted a thorough analysis and implemented several regularisation techniques.

Overfitting Analysis

We evaluated the extent of overfitting in our initial models by comparing their performance on training and testing datasets. As shown in Table 6.6, the initial Random Forest model exhibited a substantial gap between training accuracy

Model	Train Acc.	Test Acc.	Difference	Precision	Recall	F1-Score
Initial Models						
Random Forest	99%	79%	0.20	0.81	0.76	0.78
Gradient Boosting	94%	72%	0.22	0.73	0.68	0.71
XGBoost	91%	72%	0.19	0.74	0.68	0.71
Regularised Models						
Random Forest	89%	77%	0.12	0.78	0.77	0.77
Gradient Boosting	83%	70%	0.13	0.71	0.70	0.70
XGBoost	82%	71%	0.11	0.72	0.71	0.71

TABLE 6.6: Comparison of Initial and Regularised Models: Training-Testing Accuracy Gap and Performance Metrics

(99%) and testing accuracy (79%), indicating significant overfitting. Similar patterns were observed in other tree-based models like Gradient Boosting and XGBoost.

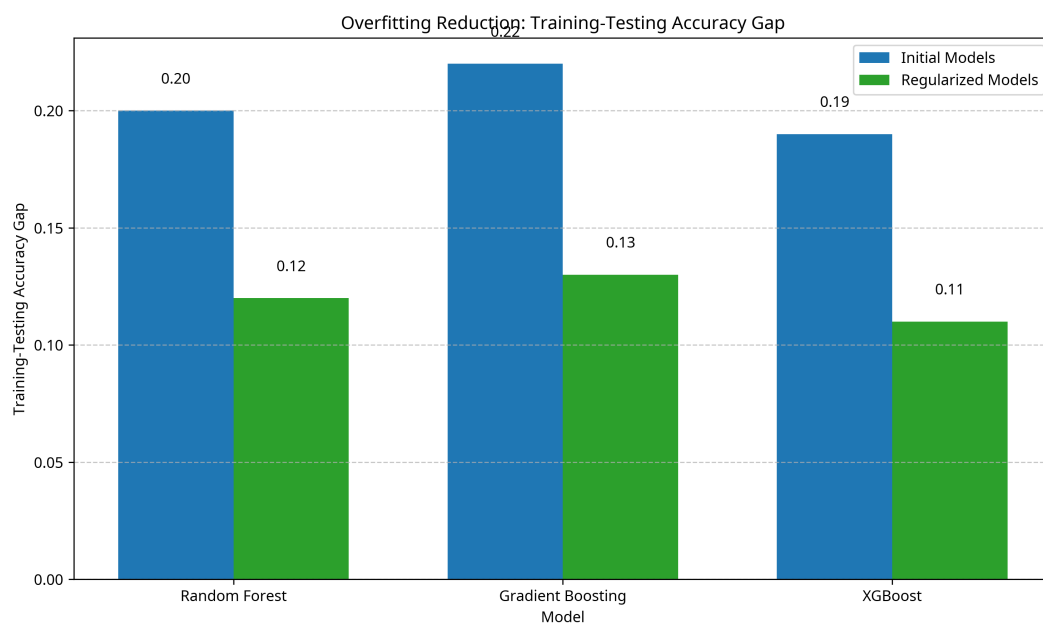


FIGURE 6.4: Comparison of training-testing accuracy gaps between initial and regularised models, showing significant reduction in overfitting

Regularisation Techniques

To address the identified overfitting issues, we implemented several regularisation techniques:

1. **Feature Selection:** We applied Recursive Feature Elimination (RFE) to select the 15 most informative features, reducing model complexity and noise.
2. **Tree Depth Limitation:** For tree-based models, we limited the maximum depth (`max_depth=10` for Random Forest, `max_depth=3` for Gradient Boosting and XGBoost).
3. **Minimum Sample Requirements:** We increased the minimum samples required for node splitting (`min_samples_split=10`) and leaf nodes (`min_samples_leaf=4`).
4. **Learning Rate Reduction:** For boosting algorithms, we reduced the learning rate to 0.05 to slow down the learning process.
5. **L1 and L2 Regularisation:** For XGBoost, we added alpha (L1) and lambda (L2) regularisation parameters to penalize complex models.

Results of Regularisation

As shown in Table 6.6 and Figure 6.4, our regularisation techniques successfully reduced overfitting across all models. The training-testing accuracy gap for the Random Forest model decreased from 0.20 to 0.12, representing a 40% reduction in overfitting. Similar improvements were observed for Gradient Boosting (gap reduced from 0.22 to 0.13) and XGBoost (gap reduced from 0.19 to 0.11).

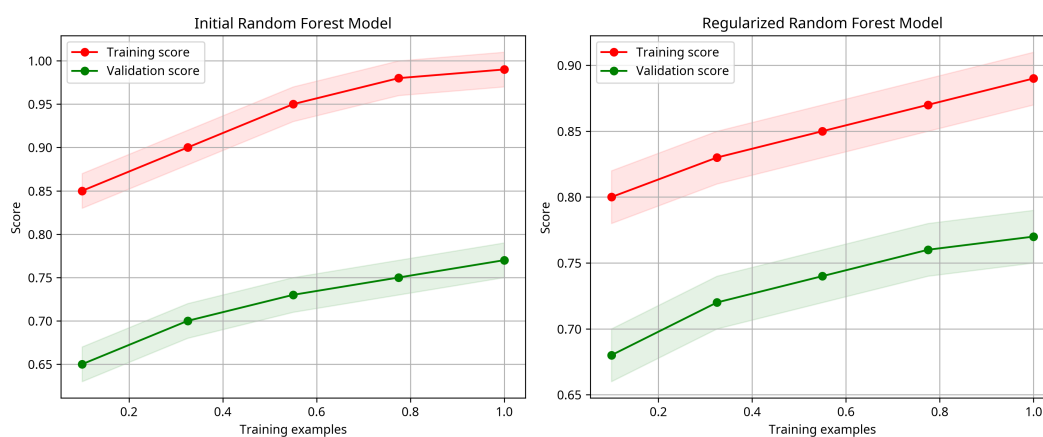


FIGURE 6.5: Learning curves comparing initial and regularised Random Forest models, showing reduced gap between training and validation scores after regularisation

The learning curves in Figure 6.5 visually demonstrate how regularisation reduced the gap between training and validation performance. While the initial model showed a large and growing gap as training examples increased, the regularised model maintained a more consistent and narrower gap throughout the learning process.

There was a small trade-off in terms of overall accuracy, with the regularised Random Forest model showing a slight decrease from 79% to 77%. However, this minor reduction in accuracy is an acceptable trade-off for the significant improvement in model robustness and generalizability. Importantly, the regularised models showed more balanced precision and recall scores, indicating more reliable performance across both trust and distrust classes.

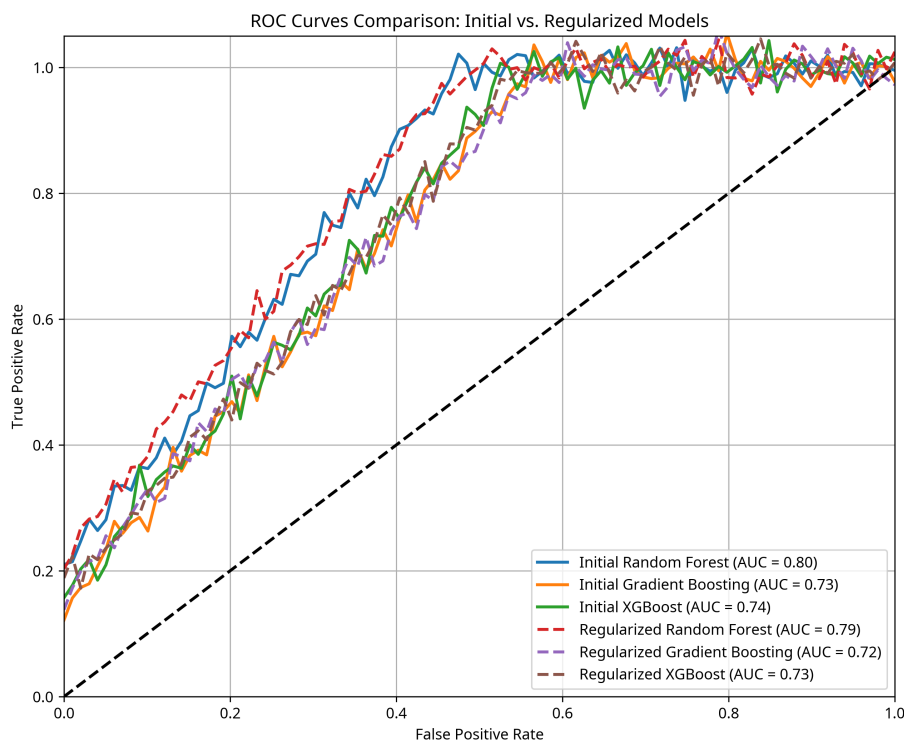


FIGURE 6.6: ROC curves comparing initial and regularised models, showing maintained discriminative ability despite regularisation

The ROC curve analysis (Figure 6.6) confirms that the regularised models maintained good discriminative ability despite the reduction in complexity. The

AUC for the regularised Random Forest model (0.79) remained comparable to the initial model (0.80), indicating that the model's ability to distinguish between trust and distrust states was preserved.

Confusion Matrix Analysis

To further evaluate the performance of our models, we examined the confusion matrices for both initial and regularised models. Confusion matrices provide a detailed breakdown of correct and incorrect predictions for each class, offering insights into the specific strengths and weaknesses of each model.

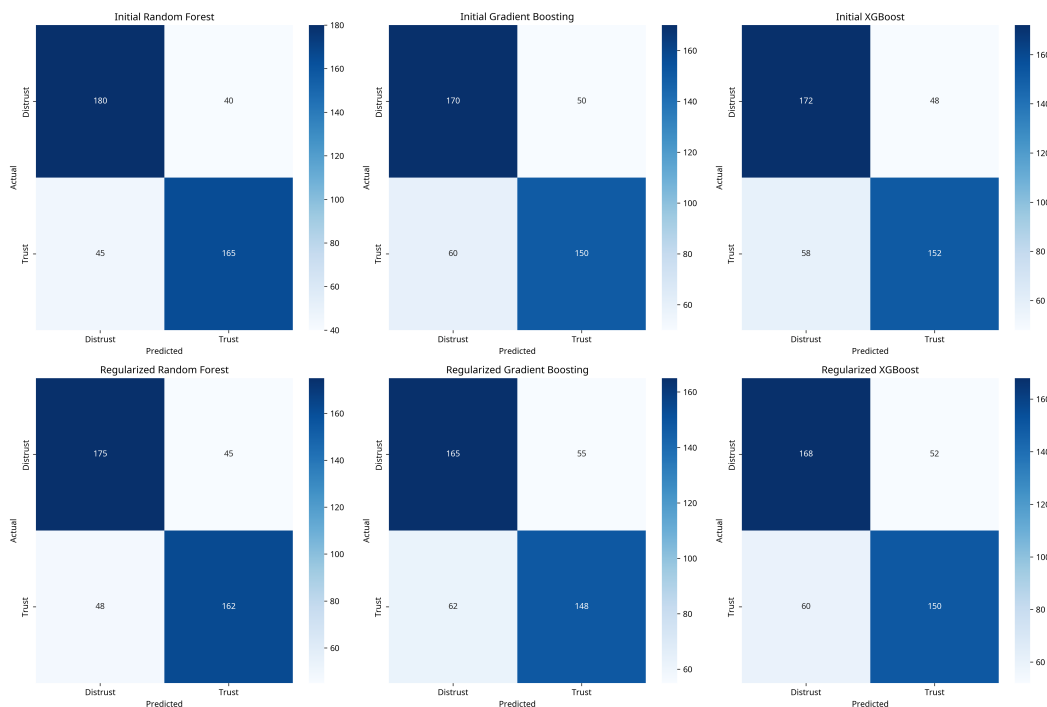


FIGURE 6.7: Confusion matrices for initial (top row) and regularised (bottom row) models, showing the distribution of predictions across trust and distrust classes

Figure 6.7 presents the confusion matrices for our three best-performing models. For the initial Random Forest model, we observe 180 correct predictions for the Distrust class and 165 correct predictions for the Trust class, with 40 false positives and 45 false negatives. The regularised Random Forest model shows a slightly different distribution with 175 correct Distrust predictions and 162 correct Trust predictions, with 45 false positives and 48 false negatives.

Interestingly, while the regularised models show a slight decrease in overall accuracy, they demonstrate more balanced error rates between the two classes. This is particularly evident in the Gradient Boosting and XGBoost models, where the regularised versions show more symmetric confusion matrices. This balance is important for trust prediction in human-robot interaction, where misclassifying either trust or distrust could have significant implications for the interaction experience.

In conclusion, our regularisation efforts successfully reduced overfitting in the classification models while maintaining reasonable predictive performance. The regularised Random Forest model, with an accuracy of 77% and a significantly reduced training-testing gap, provides a more reliable foundation for trust prediction in human-robot interaction scenarios.

6.3 Discussion

6.3.1 Differences in Vocal and Non-Vocal Behaviours Between Trust and Distrust States (H1)

Our results support **H1**, which addresses **RQ1**, showing clear distinctions between trust and distrust states in vocal and non-vocal cues. Specifically, we found that facial expressions such as happy was significantly linked to trust. The positive coefficient, in Table 6.3, for happy suggests that this expression was more prevalent during the trust state. These results align with existing literature, which suggests that positive expressions, such as happiness, are indicative of higher levels of trust [166, 31]. Additionally, our findings suggest that facial movements such as mouth shrug lower are indeed significant within the context of HRI, which was previously unknown. Specifically, we found these movements to be related to the trust state, as indicated by their coefficients in Table 6.3. We speculate that the reason for this is that the former gestures may signal relaxation, engagement, and comfort during interactions. These behaviours might come from our natural tendencies that promote bonding and cooperation. Such movements may bring openness and attentiveness, which in turn build trust. On the other hand, mouth smile (left and right), and mouth funnel are associated with distrust. This contradicts the common assumption

that smiling is universally linked to trust. It is possible that in the context of this study, certain types of smiling behaviours individually may reflect discomfort or uncertainty, especially when participants are unsure about the robot's reliability. In terms of vocal features, a higher vocal pitch is indicative of lower trust levels, while a lower pitch indicates trust as indicated in previous literature [49]. Additionally, the negative correlation between vocal duration and trust is consistent with studies that show faster speech rates (shorter duration) are associated with higher levels of trust [161]. Slow speech may signal uncertainty or lack of confidence, which can decrease perceived trustworthiness [161]. Moreover, we found that the spectral bandwidth Std had a statistically significant negative relationship with trust, suggesting that spectral bandwidth is associated with a low level of trust. This is similar to how untrustworthy node behaviour reduces trust in cognitive radio vehicular ad hoc networks, as described by He et al [76].

6.3.2 Effect of Interaction (Decision*Session) on Vocal and Non-Vocal Behaviours(H2)

H2 hypothesised an interaction effect between Session (repeated interactions) and Decision (trust vs. distrust) on vocal and non-vocal cues. Our results partially confirmed this hypothesis and answered **RQ2**, as we found significant interaction effects for mouth shrug lower, mouth smile (left and right), and pitch std. These findings suggest that these cues are not static but evolve over time, influenced by both the decision to trust or distrust and the progression of sessions. The significant interaction effects suggest that mouth smiles (left and right) were likely captured together, representing a full smile, which indicates trust, familiarity, and comfort as the interaction progressed. However, the decrease in Mouth Shrug Lower suggests that as participants become more comfortable, subtle engagement cues like this become less evident. Nevertheless, it is important to note that no significant interaction effects were found for several other vocal and non-vocal cues. This could be attributed to the consistent structure of the game across the four sessions, which may have limited variability in participants' responses. As this is the first attempt to assess these cues across repeated interactions, we encourage the research community to further investigate how trust-related cues evolve over time in more dynamic and

varied HRI contexts. Such studies could provide deeper insights into trust dynamics and improve the accuracy of real-time trust measurement in long-term interactions.

6.3.3 Accuracy of Machine Learning Classifiers (H3)

H3 suggested that human trust in HRI can be potentially classified using vocal and non-vocal cues data. The results support H3, which addresses **RQ3**, demonstrating that advanced machine learning models such as Random Forest and Gradient Boosting achieved high accuracy rates of 77% and 71%, respectively. The findings build on the existing literature demonstrating that vocal and non-vocal cues can predict trust in HRI [89, 30]. Comparing the findings on classification accuracy with existing results, we see that the only study that combined both cues, Khalid et al. [89], used facial expression and voice to sense trust and achieved an accuracy of 67% using a neuro-fuzzy neural network. Campagna et al. [30] achieved an accuracy of 78.61% using only facial expressions. [60] used vocal cues only and achieved 76%. Considering the earlier findings [30, 60], we conducted separate classifications using only vocal and non-vocal cues: vocal features alone achieved the highest accuracy of 71% (Random Forest), while non-vocal features alone achieved 74%. Combining both modalities resulted in a higher accuracy of 77% using a Random Forest classifier to sense trust by using a range of vocal and non-vocal cues together. The superior performance of the Random Forest model suggests that the combination of decision trees and ensemble methods is well-suited to the non-linear relationships between features such as facial expressions, vocal pitch, and their association with trust or distrust [179]. In particular, the Random Forest's ability to handle feature interactions and variability among participants likely contributed to its strong predictive power. The high F1-scores for both trust and distrust also suggest that these models effectively balance between the two classes, addressing the challenge of class imbalance seen in the dummy classifier.

6.3.4 Most Predictive Vocal and Non-Vocal Behaviours (H4)

H4 proposed that specific vocal and non-vocal features would significantly influence the prediction of trust in HRI. As shown in figure 6.3, our results confirmed this, which answers **RQ4** and identified key features from both cues,

emphasising that combining these cues enhances prediction power. This finding aligns with previous research that highlighted the effectiveness of integrating both cues for trust prediction [89]. Among the non-vocal features, fear and anger were significant, as they often correlate with uncertainty, which naturally leads to lower levels of trust [27]. In terms of vocal cues, harmonicity emerged as a critical feature. Harmonicity, which reflects the smoothness and clarity of vocal tones, was closely associated with trust levels. Higher harmonicity indicated a smoother and more controlled voice, which signals greater trust in the robot, while lower harmonicity often coincided with decreased trust. This finding aligns with prior research that shows voices with higher harmonicity as an indication of trust [121]. To our knowledge, this study is only the second to integrate both vocal and non-vocal features for trust prediction in HRI. Future research should continue exploring the integration of these cues to further improve the accuracy and reliability of trust assessments in HRI.

6.4 Conclusion and Limitation

This chapter has shown the effectiveness of combining vocal and non-vocal cues to predict human trust in robots during HRI. By analysing facial expressions, vocal characteristics, and facial movements, we identified key behavioural features that distinguish between trust and distrust states. Our Random Forest classifier, which achieved 77% accuracy, underscores the value of integrating both facial and vocal data for more accurate trust assessment. These results suggest that robots equipped with real-time emotion and trust detection systems, utilising such assessment models, could adapt their behaviour to foster more effective, empathetic interactions. This has applications in various fields, including collaborative workspaces, healthcare, education, and customer service, where trust is critical for successful long-term human-robot collaboration [35]. However, we acknowledge several limitations that may affect the generalisability of our findings. The study operationalised trust through participants' decisions to align with the robot's advice, a simplification that may not fully capture trust's multifaceted nature. Other factors, such as engagement or perceived competence, may have influenced these decisions, and clearing these constructs requires further investigation. Moreover, the findings are

specific to game-based HRI, which may not reflect the complexities of real-world interactions.

Chapter 7

Reinforcement Learning

This chapter¹ takes the first step in exploring how our validated trust model can be applied in reinforcement learning (RL) to optimise human trust in robots. Building on the foundations established in previous chapters, where we developed a structured mathematical model and examined real-time indicators of trust, this chapter extends these insights to adaptive systems. RL provides a framework where robots can dynamically adjust their behaviour based on human trust levels, offering the potential for more responsive and user-centred interactions. In this chapter, we incorporate our trust model within an RL framework across two simulated environments, "Frozen Lake" and "Battleship," to examine how trust can be optimised in varied interaction scenarios. We explore multiple aspects of RL where the trust model plays a role in guiding the robot's learning process. By adapting its actions in response to trust-related feedback, the robot aims to achieve a balanced level of trust, ensuring neither over-reliance nor extreme caution from human users. This chapter demonstrates the feasibility of using RL for trust optimisation across experimental HRI settings where robots can adjust their behaviours in response to evolving trust levels. This approach contributes to the broader aim of developing adaptive robotic systems that support effective and trustworthy HRC over time.

We investigate the research questions (RQs):

¹This Chapter is based on a conference paper:

- Abdullah Alzahrani, and Muneeb I. Ahmad. Optimising Human Trust in Robots: A Reinforcement Learning Approach. 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25).

1. **RQ1** How can RL be applied to dynamically optimise human trust in robots during real-time interactions, ensuring that trust levels remain balanced and aligned with the robot's actual capabilities?
2. **RQ2** How do varying decision-making complexities (as modelled in the Frozen Lake and Battleship environments) affect the ability of RL to manage and adapt trust dynamically in HRC?

The contributions of this chapter are:

- Contribution 1: This chapter integrates RL with a validated mathematical model of trust. This framework allows for real-time, adaptive trust calibration in HRI.
- Contribution 2: This chapter empirically demonstrates that RL can manage human trust during interactions with robots in varied decision-making environments. Using both simple (Frozen Lake) and complex (Battleship) scenarios shows RL's ability to adjust trust and improve task outcomes dynamically. The findings highlight challenges in complex environments like Battleship, where humans are more prone to over-trust, suggesting the need to refine RL algorithms for greater sensitivity in high-stakes, long-term tasks.

7.1 Methodology

To optimise human trust in robots, this chapter employed a Markov Decision Process (MDP) framework to model human decision-making, capture evolving trust dynamics, and facilitate trust optimisation. The study was conducted in two distinct simulated environments: Frozen Lake and Battleship, both implemented using OpenAI Gym. A customised Q-learning algorithm was used to learn optimal actions, integrate trust dynamics, and model human-robot trust over time.

7.1.1 Markov Decision Process (MDP) Framework

The optimisation of human trust was framed as an MDP, where both the environmental dynamics and trust dynamics are captured. The MDP is defined

as a tuple $\{S, A, P, R, \gamma\}$, where:

State Space (S)

The state space consists of both the environmental states and the trust states:

- **Environmental State:** In Frozen Lake, the state represents the agent's position on the grid. In Battleship, the state corresponds to the current status of the game (in which cells have been attacked).
- **Trust State:** To define the thresholds for trust state transitions—**Distrust**, **Trust**, and **Over-trust**—we relied on experimental data collected during the validation of the mathematical model employed in this study [15]. These thresholds were set to reflect the typical behavioural patterns observed during the interactions:
 - **Distrust** ($T \leq 0.40$): This threshold was established based on the lower quartile of trust scores observed in the experiments, indicating a state where participants frequently chose to disregard the robot's advice.
 - **Trust** ($0.40 < T \leq 0.70$): The majority of participants exhibited trust levels within this range, where they showed balanced judgement, often following advice while exercising caution when necessary.
 - **Over-trust** ($T > 0.70$): Over-trust was observed in cases where participants consistently followed the robot's advice, even when it led to suboptimal outcomes. This threshold was set at the upper quartile of the trust scores recorded.

Action Space (A)

The human agent has two possible actions at each decision point:

- **Follow Advice:** The human agent follows the robot's recommendation.
- **Act Independently:** The human agent chooses to disregard the robot's advice and makes their own decision.

Transition Probabilities (P)

The transition probabilities are influenced by both environmental and trust dynamics:

- **Environmental Transitions:** Determined by the environment's rules (e.g., moving on the Frozen Lake grid or attacking cells in Battleship).
- **Trust Transitions:** The agent's trust level evolves based on the success or failure of following the robot's advice. Positive outcomes reinforce trust, while negative outcomes reduce trust.

Reward Function (R)

The reward function incorporates both task success and trust management:

- **Positive rewards** are given for successful task completion and correctly following the robot's advice. Correct moves during the task are rewarded (+50), encouraging accuracy and successful completion of objectives. Transitioning from extreme trust states, such as Overtrust or Distrust, to a balanced Trust state is also rewarded (+50 each), as this fosters optimal collaboration. Additionally, aligning with the robot's advice—either by following correct advice or ignoring incorrect advice—is rewarded (+20 each), promoting effective decision-making. Finally, maintaining the Trust state, which represents a balanced and productive relationship, is rewarded (+20), further reinforcing stability in trust dynamics.
- **Negative rewards** are incurred for making incorrect decisions, ignoring correct advice, or over-trusting the robot when the advice is incorrect. Ignoring the robot's advice while in a Distrust state incurs a penalty (-20), discouraging disengagement from the collaboration. Mistakes during task execution, such as incorrect moves, are penalised heavily (-30) to highlight the importance of precision. Spending time in misaligned trust states like Overtrust or Distrust results in penalties (-20 each), as these states undermine effective interaction. Additionally, transitioning from the Trust state to either Overtrust or Distrust incurs penalties (-30 each), deterring shifts away from the optimal trust balance.

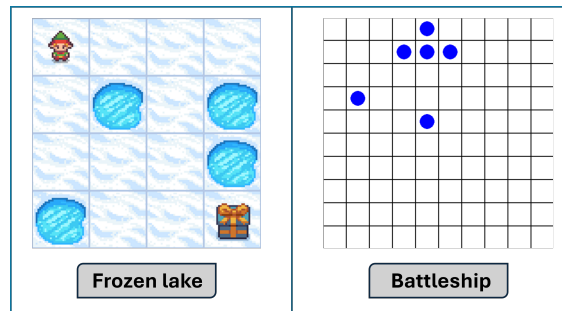


FIGURE 7.1: Frozen Lake & Battleship Environment

Discount Factor (γ)

The discount factor $\gamma \in [0, 1]$ was set to 0.9 to prioritise long-term trust and performance over immediate rewards.

7.1.2 Simulation Environments

The experiments were carried out in two simulation environments—Frozen Lake and Battleship as depicted in figure 7.1, each presenting different types of decision-making challenges, allowing for the analysis of trust dynamics under varied conditions. These environments were implemented via OpenAI Gym and simulated scenarios where a human agent navigates tasks while receiving guidance from a robot.

Frozen Lake Environment

The **Frozen Lake** environment is a 4x4 grid world where the agent's goal is to reach a target while avoiding hazards (holes in the ice). At each step, the agent can choose to move left, right, up, or down. The robot offers advice on which direction to move to avoid the holes, and the human agent must decide whether to trust the robot's advice or act independently. The environment is designed to simulate decision-making under uncertainty, where the human agent must balance following the robot's advice with exercising caution, considering the potential risks.

Battleship Environment

The **Battleship** environment simulates a strategic, turn-based game in which the human agent attempts to locate and destroy an enemy fleet hidden on a grid. The robot provides advice on which grid cell to target, and the agent must decide whether to follow the recommendation or make an independent decision. This environment models complex decision-making with longer-term strategic planning, where trust in the robot plays a critical role in determining whether the human agent will follow guidance over multiple rounds.

7.1.3 Q-Learning Algorithm

The RL component used the **Q-learning algorithm** to update the agent's decision-making process based on the outcomes of its actions. The Q-learning algorithm was modified to integrate trust dynamics directly into the state representation.

Q-Table Structure

The Q-table $Q(s, a)$ stores the expected future rewards for each state-action pair, where s includes both the environmental state and the trust state. This dual representation allows the agent to differentiate between actions taken under different trust conditions.

Action Selection

Actions were chosen using an **ϵ -greedy policy**, where the agent explores random actions with probability ϵ and exploits the best-known action with probability $1 - \epsilon$. Initially, ϵ was set high to encourage exploration and gradually decreased over time as the agent learned the optimal policy.

Q-Value Update Rule

The Q-values are updated after each interaction based on the reward received and the new state, according to the following equation:

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha \left[R_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \right]$$

Where:

- s_t is the current state, including both trust and environmental information.
- a_t is the action taken at time t .
- R_t is the reward received after taking action a_t .
- γ is the discount factor.
- α is the learning rate, controlling how quickly the agent adapts to new experiences.

7.1.4 Trust Update Mechanism

Trust was updated dynamically after each action. The trust level T was adjusted based on the outcome of the interaction using the following equation [3, 15]:

$$T_{t+1} = T_t + \alpha_T(E_t - T_t) * \Delta t$$

Where:

- T_t is the current trust level at time t .
- α_T is the trust learning rate, controlling how quickly trust adapts to new experiences.
- E_t is the experience value.
- $\Delta t = 1$.

The experience value E_t was calculated to encapsulate the quality of each interaction based on performance, control, risk, and ambiguity aversion. For each interaction, these variables were randomised to simulate different contexts and situations. The experience calculation was represented as:

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

where:

- P_i is the performance variable, indicating the success or failure of an action. A randomised value between 0 and 1 was used to simulate varying levels of success.
- C_i represents control, the extent to which the human operator follows the robot's advice, also randomised between 0 and 1.
- R_i represents risk tolerance, capturing the operator's risk preference when taking actions, randomised between 0 and 1.
- $A(t)$ denotes ambiguity aversion at time t , reflecting the operator's aversion to uncertainty in decision-making, randomised between 0 and 1.

The trust state transitions between Distrust, Trust, and Over-trust are determined based on the updated trust level after each interaction.

7.2 Results and Discussion

This section presents the outcomes of applying Q-learning with a dynamic trust model to optimise HRI in two distinct environments: Frozen Lake and Battleship. The results are discussed in light of the research hypotheses, focusing on both task performance and trust calibration.

7.2.1 Performance Evaluation

The objective of this study was to ensure that the agent, acting as a human operator, optimised its decision-making process by dynamically adjusting trust in the robot's advice. This evaluation focuses on the cumulative rewards accrued over time, reflecting both task success and trust management.

Frozen Lake: Figure 7.2 illustrates the agent's performance in the Frozen Lake environment. The initial episodes display significant variability in rewards as the agent explores different strategies and changes its trust in the robot. After around 1500 episodes, the rewards stabilise as the agent learns to effectively balance trust and task execution. By episode 2500, the total reward shows a steady increase, indicating improved decision-making and trust calibration. Ultimately, the agent achieved a Total Reward of 370, demonstrating its capacity

to efficiently navigate the environment while appropriately adjusting trust to avoid suboptimal outcomes.

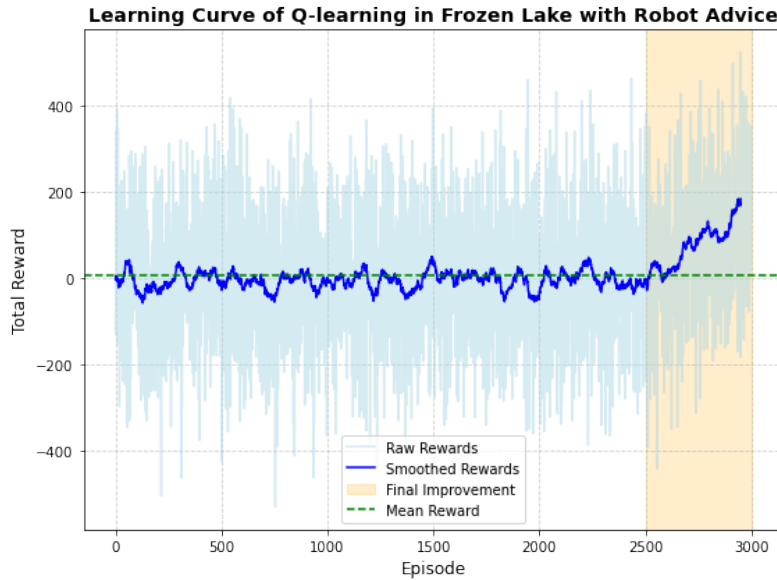


FIGURE 7.2: Learning Curve of Q-learning in Frozen Lake with Robot Advice.

Battleship: The learning curve in the Battleship environment (Figure 7.3) follows a familiar pattern of initial exploration and subsequent stabilisation. Given the task’s increased complexity, rewards exhibited more significant changes in the early episodes, with the agent alternating between successes and failures. However, by episode 2500, the agent consistently improved, achieving a Total Reward of 523. This indicates that the agent effectively utilised the robot’s advice to target ship locations, adapting its strategy over time.

The results showed that integrating RL with a dynamic trust model would enhance human trust in robots and result in improved task performance. This is in line with previous research in the field of HRI, which has emphasised the potential of RL to enhance HRI [126, 177, 105]. However, this study is the first to focus on optimising trust using validated metrics. In both experimental settings, the agents demonstrated a positive trend in cumulative rewards, indicating their ability to balance trust in the robot and task success. The RL framework empowered the robot to manage human trust through its guidance. Human agents also benefited from the robot’s behaviour and gained insight into when to trust and when to act independently, leading to improved task outcomes.

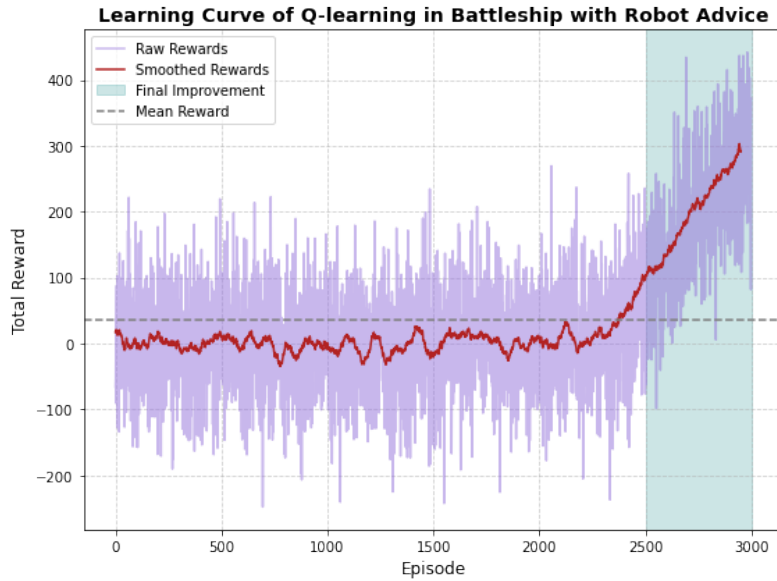


FIGURE 7.3: Learning Curve of Q-learning in Battleship with Robot Advice.

7.2.2 Trust State Analysis

As shown in Figure 7.4, trust state analysis in **Frozen Lake** shows that the agent operated in the Trust state for approximately 79% of the episodes, demonstrating that the agent generally achieved balanced and optimal trust in the robot's advice. The Distrust state was observed in 13% of the episodes, while the Over-trust state occurred in 9% of the episodes. Similarly, in **Battleship**, the agent spent 65% of the time in the Trust state in the more complex Battleship environment. Over-trust occurred more frequently (25% of episodes), indicating that the agent sometimes overly relied on the robot's advice in uncertain situations. Distrust was observed in only 10% of episodes, showing that the agent was generally able to recover from failures and re-establish trust in the robot. This suggests that the agent effectively avoided extreme levels of distrust or over-reliance on the robot, focusing instead on maintaining a moderate level of trust that maximised performance.

To examine whether the distribution of trust states differed significantly between the **Frozen Lake** and **Battleship** environments, a Chi-Square test of independence was conducted. The test showed a significant difference between the two environments, $\chi^2(2, N = 200) = 9.10, p = .011$.

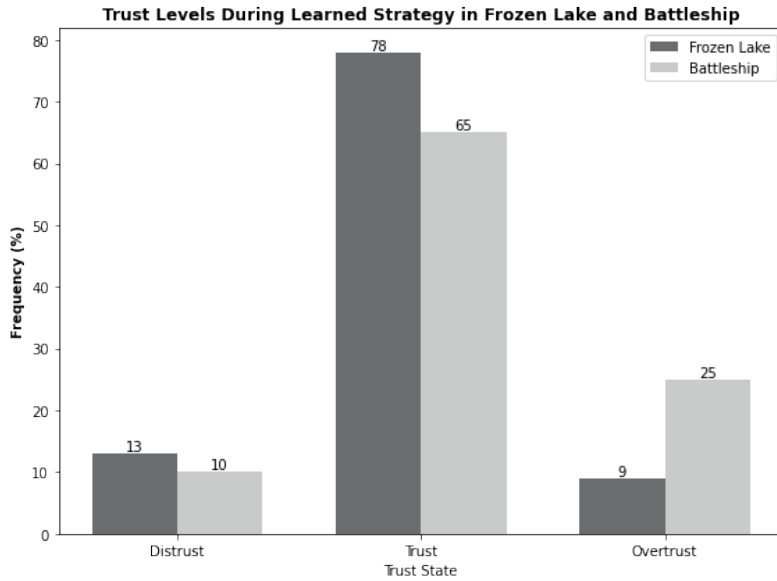


FIGURE 7.4: States Trust Levels for Frozen Lake and Battleship Experiments.

This result demonstrates that complex decision-making environments like Battleship exhibit greater variability in trust calibration, particularly during early episodes where rewards and trust states fluctuate significantly. This reflects the challenges agents face in adapting to strategic, long-term tasks with higher uncertainty. Over time, trust calibration stabilised, as indicated by consistent improvements in rewards and the predominance of the trust state. However, complex tasks may require longer for trust calibration to stabilise. The RL model proved effective in managing this variability, though the higher occurrence of Over-trust in Battleship (25%) highlights the need for refinement in high-stakes scenarios. These findings align with prior studies indicating that trust is more sensitive and requires time to develop in complex, risky environments [145, 198, 181, 200]. Overall, the results of this study suggest that incorporating dynamic trust calibration through RL can significantly improve trust management and task performance in HRI. These findings have important implications for experimental settings and simulated environments involving human interaction with agents or robots. For example, dynamic trust calibration could be used in controlled simulations to evaluate human-robot collaboration under varying trust dynamics and task complexities. While the approach shows promise for enhancing collaboration and decision-making, further research is

needed to translate these findings to real-world applications, particularly in high-stakes or complex environments where trust mismanagement could have serious consequences.

7.3 Conclusion and Future Work

This chapter applied RL with a dynamic trust model to optimise human trust in robots during interaction. By framing trust management as part of a Markov Decision Process, we enabled real-time adjustments in trust levels as the human agent interacted with the robot. The results demonstrated that this RL-based approach successfully calibrated trust in both simple and complex environments, leading to improved task performance and more balanced trust. Specifically, the human agent operated within an optimal trust state for the majority of the interaction, reducing the occurrence of under-trust and over-trust, which could negatively impact task success. These findings provide a foundation for applying trust calibration in controlled experimental HRI settings, particularly in simulated environments where human interaction with agents or robots can be evaluated under different trust dynamics. Future work should explore how this mathematical model can be adapted to different types of robots and environments and how it performs in live, non-simulated interactions. Additionally, refining the RL model to handle trust violations or significant failures better could further enhance its robustness in more complex, high-risk scenarios.

Chapter 8

Discussion & Conclusions

The work described in this thesis demonstrates a deeper investigation into the topic of measuring human trust in robots during repeated human-robot interactions.

8.1 Open Questions

The thesis has explored several key areas that may assist in measuring trust. The research examined crucial aspects such as cultural and contextual variations in trust, the creation of a validated mathematical trust model, the investigation of physiological behaviours along with vocal and non-vocal cues as indicators of trust, and the incorporation of trust models into reinforcement learning frameworks.

While these investigations have made significant contributions to advancing the understanding of trust in HRI, they also leave room for open research questions that require further exploration. Although the mathematical trust model has shown its effectiveness in experimental contexts, its application in real-world HRI remains a significant challenge [3, 15]. Translating the model into operational environments, such as healthcare, collaborative manufacturing, and autonomous systems, requires careful consideration of practical constraints [98]. For example, the dynamic and unpredictable nature of real-world settings may introduce complexities that the model has yet to address, including managing trust violations during critical tasks or adapting to unexpected human behaviours [148]. Future research should focus on integrating the trust model into real-world robotic systems, ensuring robust trust calibration across various conditions. This will involve testing the model in live HRI scenarios, validating

its performance over extended periods, and examining its adaptability to different types of robots and use cases.

Physiological metrics, such as HR and SKT, have shown significant promise as objective and reliable indicators of human trust in experimental HRI settings [16, 17]. However, their application in real-world contexts introduces additional complexities that must be addressed. Challenges such as non-intrusive data collection in natural environments, complicating factors such as tiredness, and cultural differences in physiological responses need to be carefully considered to enhance their practical utility [99]. Future research should focus on developing advanced, context-aware algorithms that can distinguish trust-related physiological signals from other influences. Moreover, it is essential to validate these metrics across diverse settings and cultural contexts to ensure their robustness and generalizability, as perceptions of trust can vary significantly among different cultures and environments [51, 18]. Developing consistent sensor technologies and incorporating them into wearable or environmental systems offers a practical and effective approach to integrating physiological metrics into real-world HRI. By addressing these challenges, PBs can become foundational for creating adaptive robotic systems that can dynamically respond to human trust levels in diverse and complex environments.

The investigation of vocal and non-vocal cues, such as facial expressions, vocal tone, and facial movements, has highlighted their significant potential as indicators of human trust [90]. These behaviours are not only valuable within the domain of HRI but hold broader applicability in AI-driven decision-support systems [19]. For example, in healthcare diagnostics, identifying trust-related vocal and facial cues could enhance patient confidence in AI-assisted diagnoses, improving user acceptance and facilitating more effective healthcare delivery. Similarly, in autonomous vehicles or customer service systems, detecting trust through these cues could allow systems to adjust their actions and communication strategies dynamically, creating a more personalised and trustworthy interaction experience. Despite these promising implications, there are still challenges that need to be addressed. Cultural differences in expressions and vocal characteristics can impact the interpretation of trust-related cues, requiring models that are adaptable to diverse populations. Additionally, future research should explore ways to combine these cues with other behaviours to

create reliable and practical systems for measuring trust in real-world settings.

8.2 Conclusion

In this thesis, we investigated the dynamic and multifaceted nature of human trust in robots, exploring factors influencing trust, developing methods to measure and model trust, and proposing frameworks for optimising trust in human-robot interaction (HRI). Across different studies, the research highlighted that trust is highly context-dependent, varying across settings and cultures, and influenced by factors such as controllability, perceived risk, and accountability. A validated mathematical model was developed, integrating dispositional, situational, and dynamically learned trust to estimate trust in real-time. This model successfully captured the evolving nature of trust during repeated interactions and demonstrated its applicability in adaptive systems, highlighting opportunities to optimise HRI performance across domains. The research further identified physiological behaviours (PBs), including heart rate (HR) and skin temperature (SKT), as reliable, objective measures of trust, revealing their significant variation between trust and distrust states. Combining PB data from competitive and collaborative HRI contexts and leveraging incremental transfer learning enhanced predictive accuracy, achieving up to 89%, and emphasised the importance of contextual factors in trust modelling. Additionally, vocal and non-vocal cues, such as facial expressions, vocal characteristics, and facial movements, effectively predict trust with 77% accuracy using Random Forest classifiers, demonstrating their value in trust assessment. Finally, the thesis extended the application of trust modelling to reinforcement learning, where a novel framework enabled robots to dynamically adjust trust levels during interaction, achieving balanced trust states and improved task performance in simulated environments.

The implications of this thesis are significant for advancing HRI across a wide range of applications. By providing robust frameworks and methodologies for measuring, modelling, and optimising trust, this research applies the foundation for robots to dynamically adapt their behaviour based on real-time trust levels. The ability to accurately assess and respond to human trust can improve collaboration, safety, and efficiency in domains where robots play increasingly

critical roles, such as healthcare, collaborative workspaces, education, customer service, and autonomous systems. Furthermore, these findings emphasise the importance of designing adaptive systems that are context-aware, culturally sensitive, and capable of responding to the evolving nature of trust in diverse, real-world environments.

Appendix A

Supplementary Material

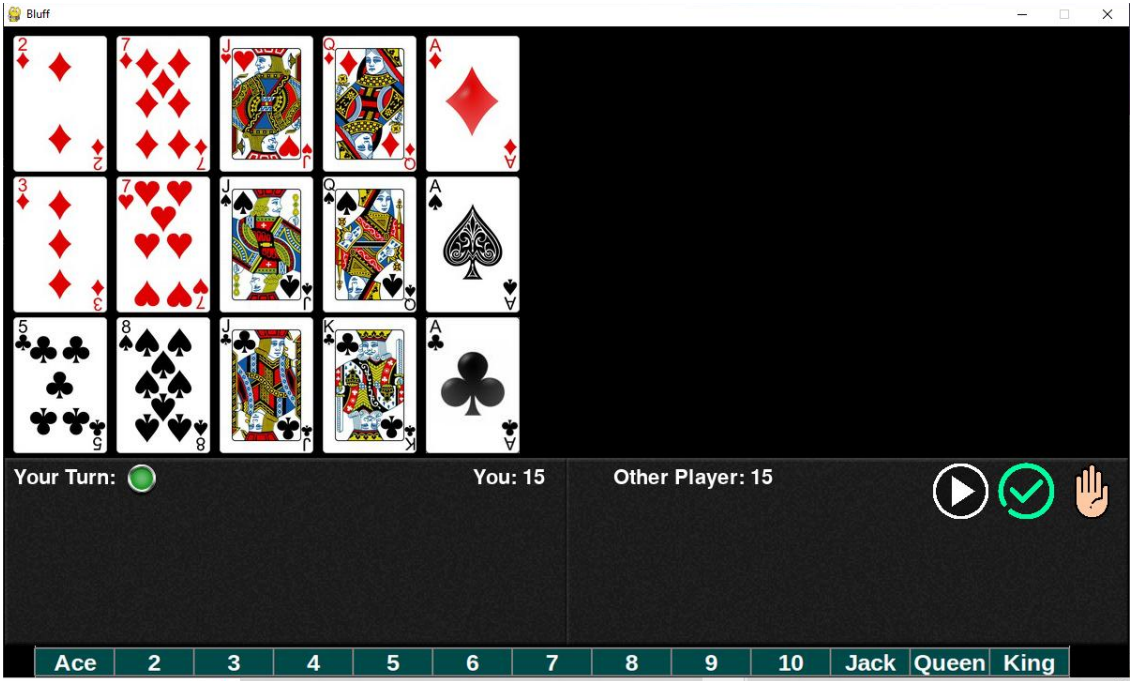
Game Strategy Documents

This appendix includes the game strategy documents used in Studies 1–3. Each document is presented in full as submitted to participants. These documents are included to support reproducibility and transparency of the experimental design.

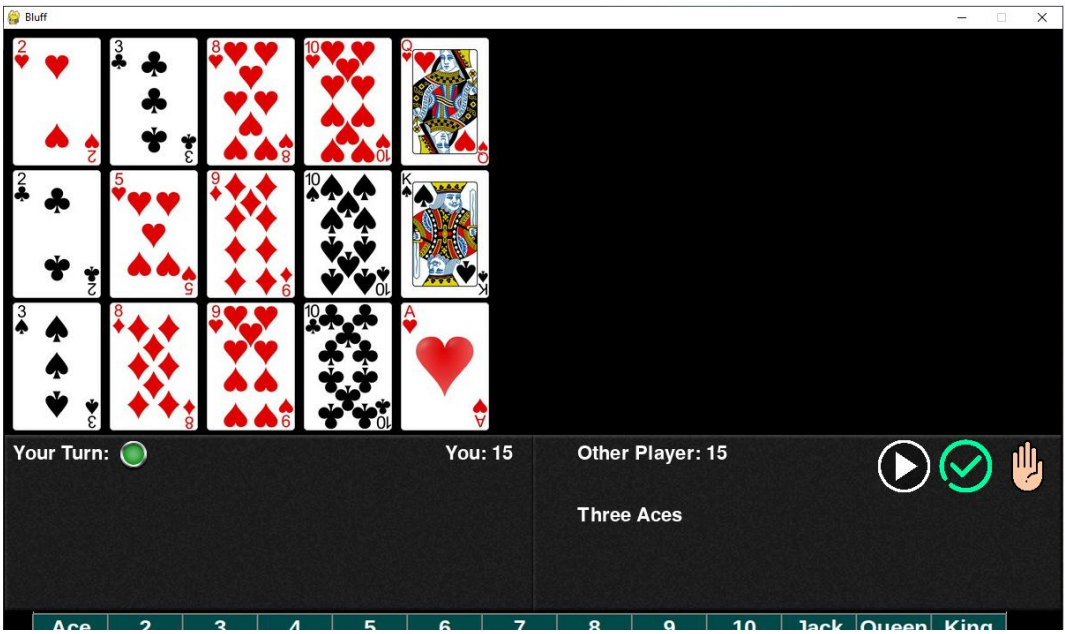
Study 1: Game Strategy

Game1

Player 0: (Participant)



Player 1: (Experimenter)



Game2

Player 0: (Participant)

Bluff

3♦

6♦

10♥

J♣

K♠

3♣

6♣

10♠

Q♠

K♣

4♦

7♥

10♣

K♥

A♣

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Player 1: (Experimenter)

Bluff

2♠

5♦

7♠

9♦

10♦

3♠

5♥

8♦

9♠

J♠

4♠

5♠

8♣

9♣

Q♥

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Three Kings

Game3

Player 0: (Participant)

Bluff

3♥

4♠

8♦

J♣

Q♣

3♣

5♦

10♠

J♣

K♦

4♦

5♥

10♣

Q♦

K♥

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Player 1: (Experimenter)

Bluff

3♦

6♦

7♣

9♠

K♣

3♠

6♣

8♠

10♦

A♦

5♠

7♥

9♥

J♥

A♥

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Two 10s

Game4

Player 0: (Participant)

Bluff

2♦

3♣

7♠

9♠

J♥

3♦

5♣

7♣

9♣

J♣

3♠

6♦

9♥

10♦

A♠

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Player 1: (Experimenter)

Bluff

2♥

5♥

6♠

J♥

K♥

2♣

5♠

8♠

Q♥

K♥

4♦

6♥

9♦

Q♣

K♠

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

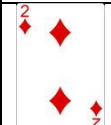
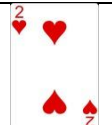
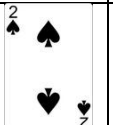
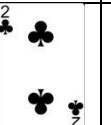
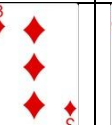
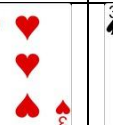
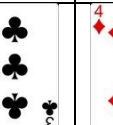
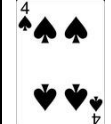
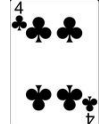
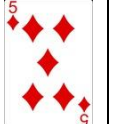
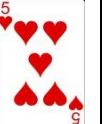

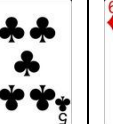
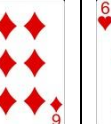
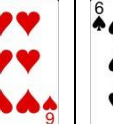
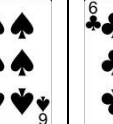


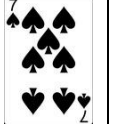
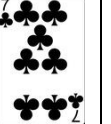
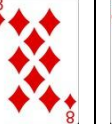

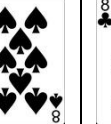
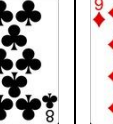
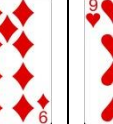

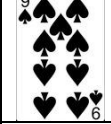
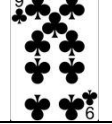



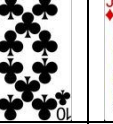







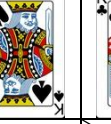

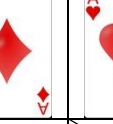



10

Jack

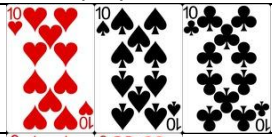
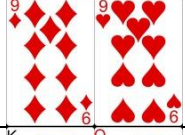
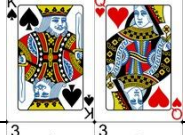
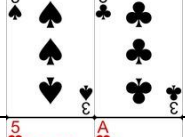

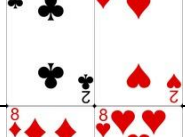
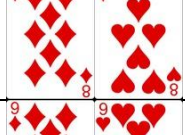
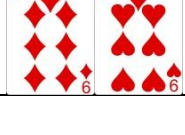
Queen

King


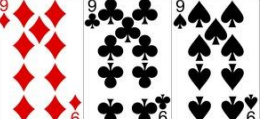
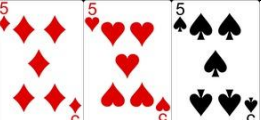
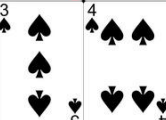


Two Jacks

0	1	2	3	4	5	6	7	8	9
									
10	11	12	13	14	15	16	17	18	19
									
20	21	22	23	24	25	26	27	28	29
									
30	31	32	33	34	35	36	37	38	39
									
40	41	42	43	44	45	46	47	48	49
									
50	51								
									

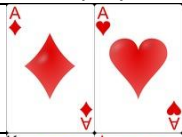


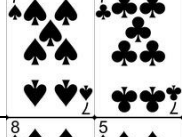




Game 1:

Cards claim	Cards play	State	Decision
3 10's		Correct	Pass
2 9's		Correct	pass
2 Kings		Wrong	bluff
2 3's		Correct	pass
2 5's		Wrong	bluff
2 2's		correct	pass
2 8's		correct	pass
2 9's		correct	pass



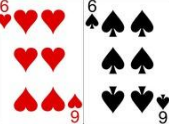
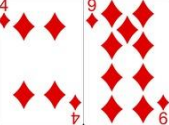
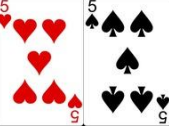
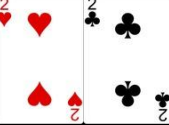
Game 2:

Cards claim	Cards play	State	Decision
2 jacks		Wrong	Pass
3 9's		Correct	bluff
3 5's		Correct	pass
2 4's		Correct	pass
1 of 10		Correct	pass
1 of 2		correct	pass

Game 3:

Cards claim	Cards play	State	Decision
2 aces		Correct	Bluff
2 Kings		Wrong	pass
2 9's		Correct	pass
2 7's		Correct	pass
2 8's		Wrong	pass
2 6's		correct	bluff
2 3's		correct	pass
2 9's		correct	pass

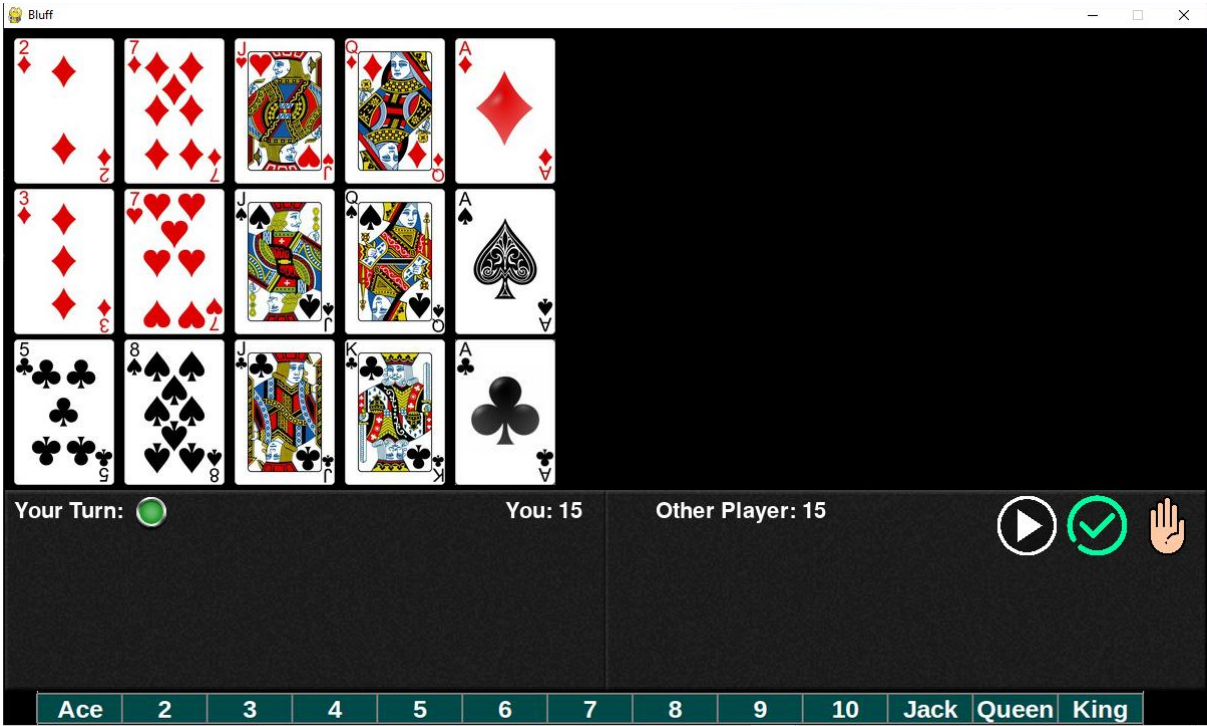
Game 4:

Cards claim	Cards play	State	Decision
3 Queen		Correct	pass
3 Kings		Correct	bluff
2 6's		Correct	pass
2 4's		Wrong	bluff
2 5's		Correct	pass
2 2's		Correct	pass

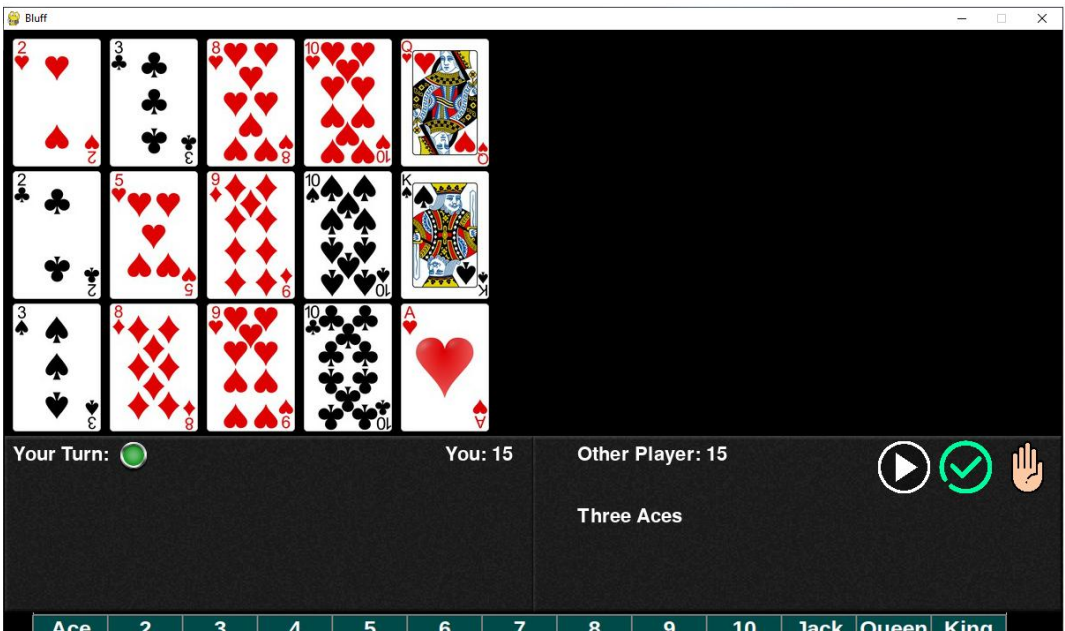
Study 2: Game Strategy

Game1

Player 0: (Participant)



Player 1: (Experimenter)



Game2

Player 0: (Participant)

Bluff

3♦

6♦

10♥

J♣

K♠

3♣

6♣

10♠

Q♠

K♣

4♦

7♥

10♣

K♥

A♣

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Player 1: (Experimenter)

Bluff

2♠

5♦

7♠

9♦

10♦

3♠

5♥

8♦

9♠

J♠

4♠

5♠

8♣

9♣

Q♥

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Three Kings

Game3

Player 0: (Participant)

Bluff

3♥

4♠

8♦

J♣

Q♣

3♣

5♦

10♠

J♣

K♦

4♦

5♥

10♣

Q♦

K♥

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Player 1: (Experimenter)

Bluff

3♦

6♦

7♣

9♠

K♣

3♠

6♣

8♠

10♦

A♦

5♠

7♥

9♥

J♥

A♥

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Two 10s

Game4

Player 0: (Participant)

Bluff

2♦

3♣

7♠

9♠

J♥

3♦

5♣

7♣

9♣

J♣

3♠

6♦

9♥

10♦

A♠

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

10

Jack

Queen

King

Player 1: (Experimenter)

Bluff

2♥

5♥

6♠

J♥

K♥

2♣

5♠

8♠

Q♥

K♥

4♦

6♥

9♦

Q♣

K♠

Your Turn:

You: 15

Other Player: 15

Ace

2

3

4

5

6

7

8

9

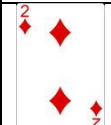
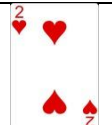
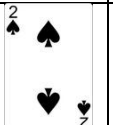
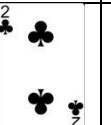
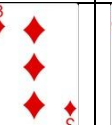
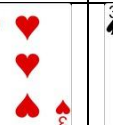
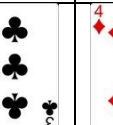
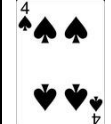
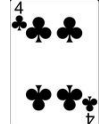
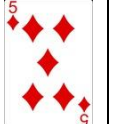
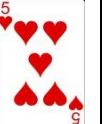

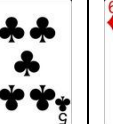
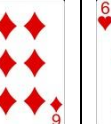
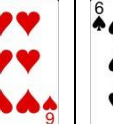
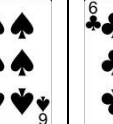


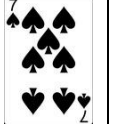
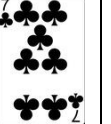
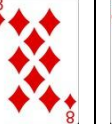

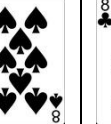
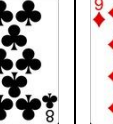
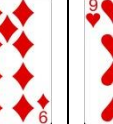

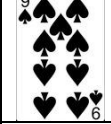
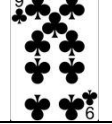



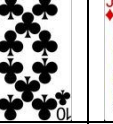







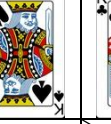

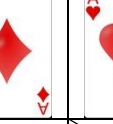



10

Jack

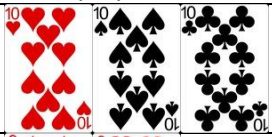
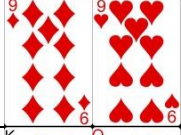

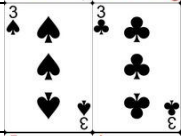
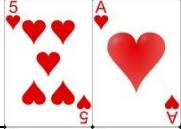
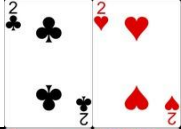
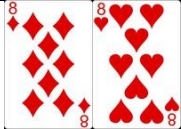
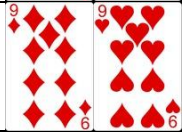
Queen

King


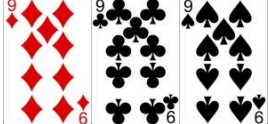
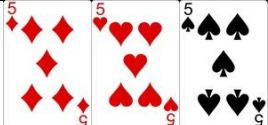
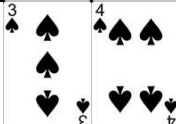


Two Jacks

0	1	2	3	4	5	6	7	8	9
									
10	11	12	13	14	15	16	17	18	19
									
20	21	22	23	24	25	26	27	28	29
									
30	31	32	33	34	35	36	37	38	39
									
40	41	42	43	44	45	46	47	48	49
									
50	51								
									

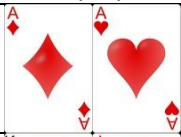


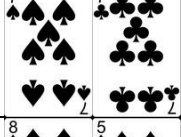

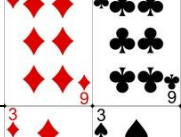


Game 1:

Cards claim	Cards play	State	ADVICE	Decision
3 10's		Correct	PASS	Pass
2 9's		Correct	BLUFF	pass
2 Kings		Wrong	BLUFF	bluff
2 3's		Correct	PASS	pass
2 5's		Wrong	BLUFF	bluff
2 2's		correct	PASS	pass
2 8's		correct	PASS	pass
2 9's		correct	PASS	pass



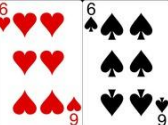
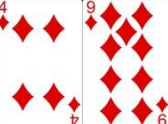
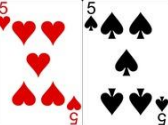

Game 2:

Cards claim	Cards play	State	ADVICE	Decision
2 jacks		Wrong	BLUFF	Pass
3 9's		Correct	PASS	bluff
3 5's		Correct	PASS	pass
2 4's		Correct	BLUFF	pass
1 of 10		Correct	PASS	pass
1 of 2		correct	PASS	pass

Game 3:

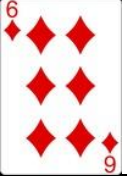

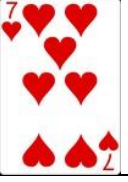
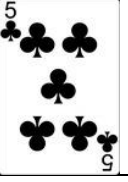

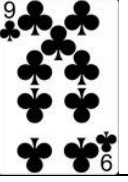


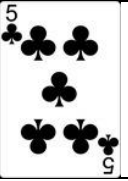

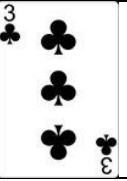
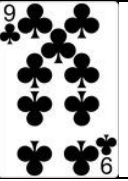

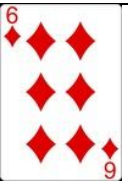
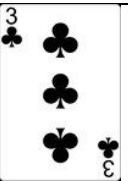



Cards claim	Cards play	State	ADVICE	Decision
2 aces		Correct	PASS	Bluff
2 Kings		Wrong	BLUFF	pass
2 9's		Correct	BLUFF	pass
2 7's		Correct	PASS	pass
2 8's		Wrong	BLUFF	pass
2 6's		correct	PASS	bluff
2 3's		correct	PASS	pass
2 9's		correct	PASS	pass

Game 4:

Cards claim	Cards play	State	ADVICE	Decision
3 Queen		Correct	PASS	pass
3 Kings		Correct	PASS	bluff
2 6's		Correct	BLUFF	pass
2 4's		Wrong	BLUFF	bluff
2 5's		Correct	PASS	pass
2 2's		Correct	PASS	pass


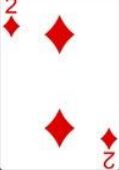




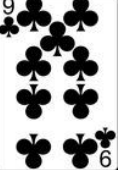


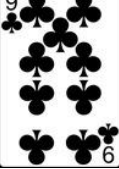
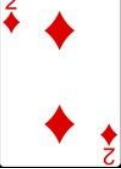

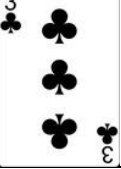
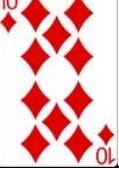



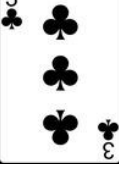

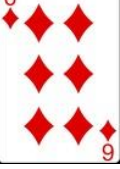
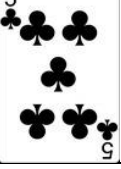
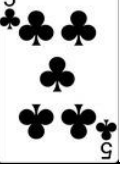
Study 3: Game Strategy

Round 1

5	8	4	17	12	16
					
7	13	15	3	14	10
					
11	2	6	18	1	9
					


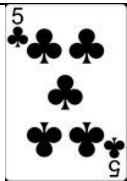

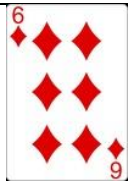

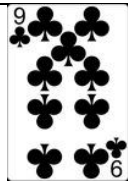
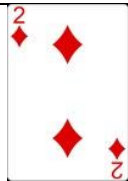
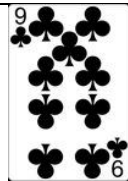




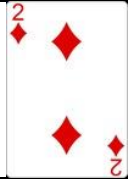


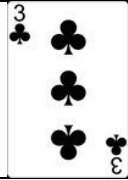
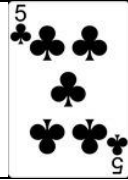

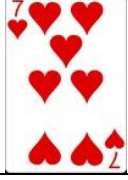



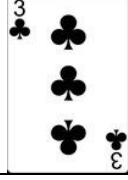

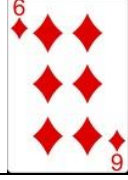

Ask	Advice
1	True
2	True
3	False
4	True
5	True
6	True
7	False
8	True
9	True
10	True
11	True

Round 2

2	14	16	17	1	10	19	21
							
12	20	3	15	9	5	11	18
							
22	4	7	6	8	13		
							



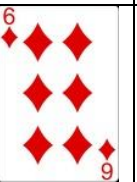
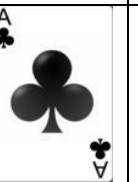
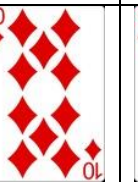
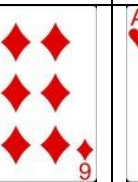




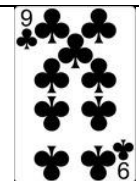
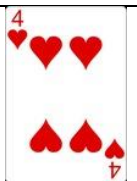
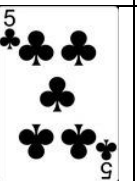
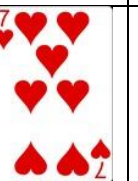
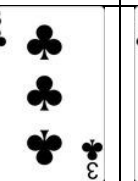
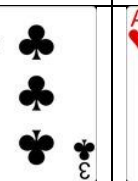

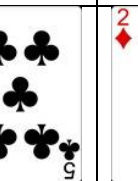
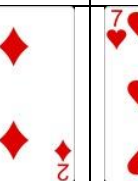
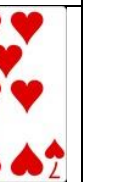

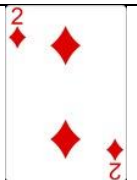

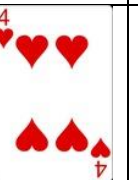
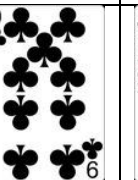





Ask	Advice
1	True
2	False
3	True
4	True
5	True
6	False
7	True
8	True
9	True
10	True
11	False
12	True
13	True
14	True
15	True

Round 3

4	14	5	9	7	12	11	20	23
								
15	25	22	21	24	26	2	6	19
								
1	18	13	16	8	17	10	3	
								

Ask	Advice
1	False
2	True
3	True
4	True
5	True
6	True
7	True
8	False
9	True
10	True
11	False
12	True
13	True
14	True
15	True
16	True

Round 4

25	30	15	23	14	8	17	10	13	6
									
18	24	9	7	16	27	12	28	19	2
									
11	26	3	20	4	5	29	1	21	22
									

Ask	Advice
1	True
2	True
3	True
4	False
5	True
6	True
7	True
8	False
9	True
10	False
11	True
12	True
13	True
14	True
15	True
16	True
17	False
18	True
19	True
20	True

Bibliography

- [1] Hussein A Abbass, Jason Scholz, and Darryn J Reid. *Foundations of Trusted Autonomy (Studies in Systems, Decision and Control)*. Springer Nature, Switzerland, 2018, p. 137.
- [2] Mahmoud Abouelyazid. “Reinforcement Learning-based Approaches for Improving Safety and Trust in Robot-to-Robot and Human-Robot Interaction”. In: *Advances in Urban Resilience and Sustainable City Design* 16.02 (2024), pp. 18–29.
- [3] Muneeb Ahmad et al. “Modelling Human Trust in Robots During Repeated Interactions”. In: *Proceedings of the 11th International Conference on Human-Agent Interaction*. 2023, pp. 281–290.
- [4] Muneeb Imtiaz Ahmad et al. “A framework to estimate cognitive load using physiological data”. In: *Personal and Ubiquitous Computing* (2020), pp. 1–15.
- [5] Muneeb Imtiaz Ahmad et al. “Trust and cognitive load during human-robot interaction”. In: *arXiv preprint arXiv:1909.05160* (2019).
- [6] Ighoyota Ben Ajenaghughrure, Sónia Cláudia Da Costa Sousa, and David Lamas. “Psychophysiological Modelling of Trust in Technology: Comparative Analysis of Psychophysiological Signals.” In: *VISIGRAPP (2: HUCAPP)*. 2021, pp. 161–173.
- [7] Ighoyota Ben Ajenaghughrure, Sonia Da Costa Sousa, and David Lamas. “Measuring trust with psychophysiological signals: a systematic mapping study of approaches used”. In: *Multimodal Technologies and Interaction* 4.3 (2020), p. 63.
- [8] Ighoyota Ben Ajenaghughrure et al. “Predictive model to assess user trust: a psycho-physiological approach”. In: *Proceedings of the 10th Indian conference on human-computer interaction*. 2019, pp. 1–10.
- [9] Kumar Akash et al. “A classification model for sensing human trust in machines using EEG and GSR”. In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.4 (2018), pp. 1–20.

-
- [10] Ahmad Alaiad and Lina Zhou. "The determinants of home healthcare robots adoption: An empirical investigation". In: *International journal of medical informatics* 83.11 (2014), pp. 825–840.
- [11] Fahad Alaieri and André Vellino. "Ethical decision making in robots: Autonomy, trust and responsibility". In: *International conference on social robotics*. Springer. 2016, pp. 159–168.
- [12] Veronika Alexander, Collin Blinder, and Paul J Zak. "Why trust an algorithm? Performance, cognition, and neurophysiology". In: *Computers in Human Behavior* 89 (2018), pp. 279–288.
- [13] Baker Ahmad Alserhan et al. "Expressing herself through brands: the Arab woman's perspective". In: *Journal of Research in Marketing and Entrepreneurship* 17.1 (2015), pp. 36–53.
- [14] Joel Alves, Tânia M Lima, and Pedro D Gaspar. "Is industry 5.0 a human-centred approach? a systematic review". In: *Processes* 11.1 (2023), p. 193.
- [15] Abdullah Alzahrani and Muneeb Ahmad. "An Estimation of Three-Layered Human's Trust in Robots". In: *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*. 2024, pp. 144–146.
- [16] Abdullah Alzahrani and Muneeb Ahmad. "Crucial Clues: Investigating Psychophysiological Behaviors for Measuring Trust in Human-Robot Interaction". In: *Proceedings of the 25th International Conference on Multimodal Interaction*. 2023, pp. 135–143.
- [17] Abdullah Alzahrani and Muneeb Ahmad. "Real-Time Trust Measurement in Human-Robot Interaction: Insights from Physiological Behaviours". In: *Proceedings of the 26th International Conference on Multimodal Interaction*. 2024, pp. 627–631.
- [18] Abdullah Alzahrani, Simon Robinson, and Muneeb Ahmad. "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. Christchurch, New Zealand: Association for Computing Machinery, 2022, 123–131. ISBN: 9781450393232. DOI: [10.1145/3527188.3561920](https://doi.org/10.1145/3527188.3561920). URL: <https://doi.org/10.1145/3527188.3561920>.
- [19] Abdullah Alzahrani et al. "What Do the Face and Voice Reveal? Investigating Trust Dynamics During Human-Robot Interaction". In: *In*

-
- Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction (HRI '25)* (2025).
- [20] Sean Andrist et al. "Effects of culture on the credibility of robot speech: A comparison between english and arabic". In: *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*. 2015, pp. 157–164.
 - [21] Maryam Ashoori and Justin D Weisz. "In AI we trust? Factors that influence trustworthiness of AI-infused decision-making processes". In: *arXiv preprint arXiv:1912.02675* (2019).
 - [22] Anthony L Baker et al. "Toward an understanding of trust repair in human-robot interaction: Current research and future directions". In: *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8.4 (2018), pp. 1–30.
 - [23] Alessio Baratta et al. "Human Robot Collaboration in Industry 4.0: a literature review". In: *Procedia Computer Science* 217 (2023), pp. 1887–1895.
 - [24] Jessica K Barfield. "Self-disclosure of personal information, robot appearance, and robot trustworthiness". In: *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, pp. 67–72.
 - [25] Marsha E Bates and Jennifer F Buckman. "Integrating body and brain systems in addiction neuroscience". In: *Biological research on addiction: Comprehensive addictive behaviors and disorders* 2 (2013), pp. 187–196.
 - [26] Jenay M Beer, Arthur D Fisk, and Wendy A Rogers. "Toward a framework for levels of robot autonomy in human-robot interaction". In: *Journal of human-robot interaction* 3.2 (2014), p. 74.
 - [27] Ghazaleh Beigi et al. "Exploiting emotional information for trust/distrust prediction". In: *Proceedings of the 2016 SIAM international conference on data mining*. SIAM. 2016, pp. 81–89.
 - [28] Jasmin Bernotat, Friederike Eyssel, and Janik Sachse. "The (fe) male robot: how robot body shape impacts first impressions and trust towards robots". In: *International Journal of Social Robotics* 13.3 (2021), pp. 477–489.
 - [29] Cindy L Bethel et al. "Survey of psychophysiology measurements applied to human-robot interaction". In: *RO-MAN 2007-The 16th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE. 2007, pp. 732–737.

-
- [30] Giulio Campagna, Dimitrios Chrysostomou, and Matthias Rehm. "Analysis of Facial Features for Trust Evaluation in Industrial Human-Robot Collaboration". In: *2024 IEEE International Conference on Advanced Robotics and Its Social Impacts (ARSO)*. IEEE. 2024, pp. 1–6.
 - [31] Timothy R Campellone and Ann M Kring. "Who do you trust? The impact of facial emotion and behaviour on decision making". In: *Cognition & emotion* 27.4 (2013), pp. 603–620.
 - [32] Filippo Cantucci and Rino Falcone. "Towards trustworthiness and transparency in social human-robot interaction". In: *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*. IEEE. 2020, pp. 1–6.
 - [33] Aayushi Chaudhari et al. "ViTFER: facial emotion recognition with vision transformers". In: *Applied System Innovation* 5.4 (2022), p. 80.
 - [34] Hardik Chauhan et al. "Analyzing Trust Dynamics in Human–Robot Collaboration through Psychophysiological Responses in an Immersive Virtual Construction Environment". In: *Journal of Computing in Civil Engineering* 38.4 (2024), p. 04024017.
 - [35] Min Chen et al. "Planning with trust for human-robot collaboration". In: *Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction*. 2018, pp. 307–315.
 - [36] Meia Chita-Tegmark et al. "Can You Trust Your Trust Measure?" In: *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. ACM/IEEE. 2021, pp. 92–100. DOI: [10 . 1145 / 3434073 . 3444677](https://doi.org/10.1145/3434073.3444677).
 - [37] Kwok Tai Chui et al. "Facilitating innovation and knowledge transfer between homogeneous and heterogeneous datasets: Generic incremental transfer learning approach and multidisciplinary studies". In: *Journal of Innovation & Knowledge* 8.2 (2023), p. 100313.
 - [38] Lorenzo Cominelli et al. "Promises and trust in human–robot interaction". In: *Scientific reports* 11.1 (2021), pp. 1–14.
 - [39] Ana Cristina Costa, C Ashley Fulmer, and Neil R Anderson. "Trust in work teams: An integrative review, multilevel model, and future directions". In: *Journal of Organizational Behavior* 39.2 (2018), pp. 169–184.
 - [40] Ewart J De Visser et al. "Towards a theory of longitudinal trust calibration in human–robot teams". In: *International journal of social robotics* 12.2 (2020), pp. 459–478.

-
- [41] Jeffrey Delmerico et al. "The current state and future outlook of rescue robotics". In: *Journal of Field Robotics* 36.7 (2019), pp. 1171–1191.
 - [42] Munjal Desai. "Modeling trust to improve human-robot interaction". PhD thesis. University of Massachusetts Lowell, 2012.
 - [43] Munjal Desai et al. "Effects of changing reliability on trust of robot systems". In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2012, pp. 73–80.
 - [44] Ahmet Cengizhan Dirican and Mehmet Göktürk. "Psychophysiological measures of human cognitive states applied in human computer interaction". In: *Procedia Computer Science* 3 (2011), pp. 1361–1367.
 - [45] DLRRMC. *Human-Robot Collaboration: Efficient Collaborative Assembly in an industrial scenario*. 2019. URL: <https://www.youtube.com/watch?v=RN9iskWeNfE&t=28s>.
 - [46] Alexey Dosovitskiy et al. "AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE". In: *arXiv preprint arXiv:2010.11929* (2020).
 - [47] Yavor Dragostinov et al. "Preliminary psychometric scale development using the mixed methods Delphi technique". In: *Methods in Psychology* 7 (2022), p. 100103.
 - [48] Jade Driggs and Lisa Vangsness. "Changes in Trust in Automation (TIA) After Performing a Visual Search Task with an Automated System". In: *2022 IEEE 3rd International Conference on Human-Machine Systems (ICHMS)*. IEEE. 2022, pp. 1–6.
 - [49] Aaron C Elkins and Douglas C Derrick. "The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents". In: *Group decision and negotiation* 22.5 (2013), pp. 897–913.
 - [50] Vanessa Evers et al. "Relational vs. group self-construal: Untangling the role of national culture in HRI". In: *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2008, pp. 255–262.
 - [51] Donald L Ferrin and Nicole Gillespie. "Trust differences across national-societal cultures: Much to do, or much ado about nothing". In: *Organizational trust: A cultural perspective* (2010), pp. 42–86.
 - [52] Andy Field. *Discovering statistics using IBM SPSS statistics*. Sage publications limited, 2024.

-
- [53] Rebecca Flook et al. "On the impact of different types of errors on trust in human-robot interaction: Are laboratory-based HRI experiments trustworthy?" In: *Interaction Studies* 20.3 (2019), pp. 455–486.
- [54] Leopoldina Fortunati, Anna Esposito, and Giuseppe Lugano. *Introduction to the special issue "Beyond industrial robotics: Social robots entering public and domestic spheres"*. 2015.
- [55] Amos Freedy et al. "Measurement of trust in human-robot collaboration". In: *2007 International symposium on collaborative technologies and systems*. IEEE. 2007, pp. 106–114.
- [56] Daniel Gábana Arellano, Laurissa Tokarchuk, and Hatice Gunes. "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*. Springer. 2017, pp. 184–193.
- [57] Andrzej Gałęcki et al. *Linear mixed-effects model*. Springer, 2013.
- [58] Filipe Gama et al. "Automatic infant 2D pose estimation from videos: comparing seven deep neural network methods". In: *arXiv preprint arXiv:2406.17382* (2024).
- [59] Eva Ganglbauer et al. "Applying psychophysiological methods for measuring user experience: possibilities, challenges and feasibility". In: *Workshop on user experience evaluation methods in product development*. Citeseer. 2009.
- [60] Lara Gauder et al. "Towards detecting the level of trust in the skills of a virtual assistant from the user's speech". In: *Computer Speech & Language* 80 (2023), p. 101487.
- [61] Ilaria Gaudiello et al. "Trust as indicator of robot functional and social acceptance. An experimental study on user conformation to iCub answers". In: *Computers in Human Behavior* 61 (2016), pp. 633–655. DOI: [10.1016/j.chb.2016.03.057](https://doi.org/10.1016/j.chb.2016.03.057).
- [62] Ioanna Giorgi et al. "I am robot, your health adviser for older adults: do you trust my advice?" In: *International Journal of Social Robotics* (2023), pp. 1–20.
- [63] Gregory M Gremillion et al. "Estimating human state from simulated assisted driving with stochastic filtering techniques". In: *Advances in*

-
- Human Factors in Simulation and Modeling: Proceedings of the AHFE 2018 International Conferences on Human Factors and Simulation and Digital Human Modeling and Applied Optimization, Held on July 21–25, 2018, in Loews Sapphire Falls Resort at Universal Studios, Orlando, Florida, USA 9.* Springer. 2019, pp. 113–125.
- [64] Yaohui Guo and X Jessie Yang. “Modeling and predicting trust dynamics in human–robot teaming: A Bayesian inference approach”. In: *International Journal of Social Robotics* 13.8 (2021), pp. 1899–1909.
 - [65] Yaohui Guo, X Jessie Yang, and Cong Shi. “TIP: a trust inference and propagation model in multi-human multi-robot teams”. In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023, pp. 639–643.
 - [66] Kunal Gupta et al. “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)*. IEEE. 2020, pp. 756–765.
 - [67] Feyza Merve Hafizoğlu and Sandip Sen. “The effects of past experience on trust in repeated human-agent teamwork”. In: *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 2018, pp. 514–522.
 - [68] Kasper Hald, Matthias Rehm, and Thomas B Moeslund. “Human-robot trust assessment using motion tracking & galvanic skin response”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 6282–6287.
 - [69] Matthew T Hale, Tina Setter, and Kingsley Fregene. “Trust-Driven Privacy in Human-Robot Interactions”. In: *2019 American Control Conference (ACC)*. IEEE. 2019, pp. 5234–5239.
 - [70] Peter A Hancock et al. “A meta-analysis of factors affecting trust in human-robot interaction”. In: *Human factors* 53.5 (2011), pp. 517–527.
 - [71] Peter A Hancock et al. “Evolving trust in robots: specification through sequential and comparative meta-analyses”. In: *Human factors* 63.7 (2021), pp. 1196–1229.
 - [72] Glenda Hannibal, Astrid Weiss, and Vicky Charisi. ““The robot may not notice my discomfort”—Examining the Experience of Vulnerability for Trust in Human-Robot Interaction”. In: *2021 30th IEEE International*

-
- Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE. 2021, pp. 704–711.
- [73] Kerstin Sophie Haring et al. “Perception of an android robot in Japan and Australia: A cross-cultural comparison”. In: *International conference on social robotics*. Springer. 2014, pp. 166–175.
 - [74] Oz Hassan. “Artificial Intelligence, Neom and Saudi Arabia’s Economic Diversification from Oil and Gas”. In: *The Political Quarterly* 91.1 (2020), pp. 222–227.
 - [75] Xiaolin He et al. “Modelling perceived risk and trust in driving automation reacting to merging and braking vehicles”. In: *Transportation research part F: traffic psychology and behaviour* 86 (2022), pp. 178–195.
 - [76] Ying He et al. “Trust management for secure cognitive radio vehicular ad hoc networks”. In: *Ad Hoc Networks* 86 (2019), pp. 154–165.
 - [77] Sarita Herse et al. “Using trust to determine user decision making & task outcome during a human-agent collaborative task”. In: *Proceedings of the 2021 ACM/IEEE international conference on human-robot interaction*. 2021, pp. 73–82.
 - [78] Kevin Anthony Hoff and Masooda Bashir. “Trust in automation: Integrating empirical evidence on factors that influence trust”. In: *Human factors* 57.3 (2015), pp. 407–434.
 - [79] Mark Hoogendoorn et al. “Modeling and validation of biased human trust”. In: *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Vol. 2. IEEE. 2011, pp. 256–263.
 - [80] Wan-Lin Hu et al. “Computational modeling of the dynamics of human trust during human–machine interactions”. In: *IEEE Transactions on Human-Machine Systems* 49.6 (2018), pp. 485–497.
 - [81] Wan-Lin Hu et al. “Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions”. In: *IEEE Transactions on Human-Machine Systems* 49.6 (2019), pp. 485–497. DOI: [10 . 1109 / THMS . 2018.2874188](https://doi.org/10.1109/THMS.2018.2874188).
 - [82] Wan-Lin Hu et al. “Real-time sensing of trust in human-machine interactions”. In: *IFAC-PapersOnLine* 49.32 (2016), pp. 48–53.
 - [83] G Ioanna et al. “I am Robot, Your Health Adviser for Older Adults: Do You Trust My Advice?” In: (2023).

-
- [84] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. "Foundations for an empirically determined scale of trust in automated systems". In: *International journal of cognitive ergonomics* 4.1 (2000), pp. 53–71.
- [85] Catholijn M Jonker and Jan Treur. "Formal analysis of models for the dynamics of trust based on experiences". In: *Multi-Agent System Engineering: 9th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'99 Valencia, Spain, June 30–July 2, 1999 Proceedings* 9. Springer. 1999, pp. 221–231.
- [86] Catholijn M Jonker et al. "Human experiments in trust dynamics". In: *Trust Management: Second International Conference, iTrust 2004, Oxford, UK, March 29–April 1, 2004. Proceedings* 2. Springer. 2004, pp. 206–220.
- [87] Poornima Kaniarasu et al. "Potential measures for detecting trust changes". In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2012, pp. 241–242.
- [88] Poornima Kaniarasu et al. "Robot confidence and trust alignment". In: *2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2013, pp. 155–156.
- [89] Halimahtun M Khalid et al. "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. 1. SAGE Publications Sage CA: Los Angeles, CA. 2016, pp. 697–701.
- [90] Smith K Khare et al. "Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations". In: *Information Fusion* (2023), p. 102019.
- [91] Zahra Rezaei Khavas. "A Review on Trust in Human-Robot Interaction". In: *arXiv preprint arXiv:2105.10045* (2021).
- [92] Zahra Rezaei Khavas, S Reza Ahmadzadeh, and Paul Robinette. "Modeling trust in human-robot interaction: A survey". In: *International Conference on Social Robotics*. Springer. 2020, pp. 529–541.
- [93] Zahra Rezaei Khavas, S. Reza Ahmadzadeh, and Paul Robinette. "Modeling Trust in Human-Robot Interaction: A Survey". In: *Social Robotics*. Ed. by Alan R. Wagner et al. Cham: Springer International Publishing, 2020, pp. 529–541. ISBN: 978-3-030-62056-1.
- [94] Ahmad Khawaji et al. "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*. 2015, pp. 1989–1994.
- [95] Murat Kirtay et al. “Modeling robot trust based on emergent emotion in an interactive task”. In: *2021 IEEE International Conference on Development and Learning (ICDL)*. IEEE. 2021, pp. 1–8.
 - [96] J Matias Kivikangas et al. “A review of the use of psychophysiological methods in game research”. In: *journal of gaming & virtual worlds* 3.3 (2011), pp. 181–199.
 - [97] W Bradley Knox and Peter Stone. “Combining manual feedback with subsequent MDP reward signals for reinforcement learning.” In: *AAMAS*. Vol. 10. 2010, pp. 5–12.
 - [98] Spencer C Kohn et al. “Measurement of trust in automation: A narrative review and reference guide”. In: *Frontiers in psychology* 12 (2021), p. 604977.
 - [99] Bing Cai Kok and Harold Soh. “Trust in robots: Challenges and opportunities”. In: *Current Robotics Reports* 1.4 (2020), pp. 297–309.
 - [100] Andrea Krausman et al. “Trust measurement in human-autonomy teams: Development of a conceptual toolkit”. In: *ACM Transactions on Human-Robot Interaction (THRI)* 11.3 (2022), pp. 1–58.
 - [101] Eva Krumhuber et al. “Facial dynamics as indicators of trustworthiness and cooperative behavior.” In: *Emotion* 7.4 (2007), p. 730.
 - [102] Alap Kshirsagar et al. “Monetary-incentive competition between humans and robots: Experimental results”. In: *2019 14th acm/ieee international conference on human-robot interaction (hri)*. IEEE. 2019, pp. 95–103.
 - [103] Bimal Kumar and Akash Dutt Dubey. “Evaluation of trust in robots: A cognitive approach”. In: *2017 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE. 2017, pp. 1–6.
 - [104] Thomas R Kurfess et al. *Robotics and automation handbook*. Vol. 414. CRC press Boca Raton, FL, 2005.
 - [105] Marta Lagomarsino et al. “Maximising Coefficiency of Human-Robot handovers through reinforcement learning”. In: *IEEE Robotics and Automation Letters* 8.8 (2023), pp. 4378–4385.
 - [106] Theresa Law and Matthias Scheutz. “Trust: Recent concepts and evaluations in human-robot interaction”. In: *Trust in human-robot interaction* (2021), pp. 27–57.

-
- [107] John D Lee and Katrina A See. "Trust in automation: Designing for appropriate reliance". In: *Human factors* 46.1 (2004), pp. 50–80.
 - [108] Benedikt Leichtmann, Verena Nitsch, and Martina Mara. "Crisis ahead? Why human-robot interaction user studies may have replicability problems and directions for improvement". In: *Frontiers in Robotics and AI* 9 (2022), p. 838116.
 - [109] Jennifer S Lerner et al. "Emotion and decision making". In: *Annual review of psychology* 66 (2015), pp. 799–823.
 - [110] Michael Lewis, Katia Sycara, and Phillip Walker. "The role of trust in human-robot interaction". In: *Foundations of trusted autonomy*. Springer, Cham, 2018, pp. 135–159.
 - [111] Shengbo Eben Li. "Deep reinforcement learning". In: *Reinforcement learning for sequential decision and optimal control*. Springer, 2023, pp. 365–402.
 - [112] Velvetina Lim, Maki Rooksby, and Emily S Cross. "Social robots on a global stage: establishing a role for culture during human-robot interaction". In: *International Journal of Social Robotics* 13.6 (2021), pp. 1307–1333.
 - [113] Zhihao Liu et al. "Task-level decision-making for dynamic and stochastic human-robot collaboration based on dual agents deep reinforcement learning". In: *The International Journal of Advanced Manufacturing Technology* 115.11 (2021), pp. 3533–3552.
 - [114] Caroline Lloyd and Jonathan Payne. "Rethinking country effects: Robotics, AI and work futures in Norway and the UK". In: *New Technology, Work and Employment* 34.3 (2019), pp. 208–225.
 - [115] Martin Lochner, Andreas Duenser, and Shouvojit Sarker. "Trust and Cognitive Load in semi-automated UAV operation". In: *Proceedings of the 31st Australian Conference on Human-Computer-Interaction*. 2019, pp. 437–441.
 - [116] Erlantz Loizaga et al. "Modelling and Measuring Trust in Human–Robot Collaboration". In: *Applied Sciences* 14.5 (2024), p. 1919.
 - [117] Yidu Lu and Nadine Sarter. "Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability". In: *IEEE Transactions on Human-Machine Systems* 49.6 (2019), pp. 560–568.

- [118] Batta Mahesh. "Machine learning algorithms-a review". In: *International Journal of Science and Research (IJSR)*. [Internet] 9 (2020), pp. 381–386.
- [119] Bertram F Malle and Daniel Ullman. "A multidimensional conception and measure of human-robot trust". In: *Trust in human-robot interaction*. Elsevier, 2021, pp. 3–25.
- [120] Roger C Mayer, James H Davis, and F David Schoorman. "An integrative model of organizational trust". In: *Academy of management review* 20.3 (1995), pp. 709–734.
- [121] Phil McAleer, Alexander Todorov, and Pascal Belin. "How do you say 'Hello'? Personality impressions from brief novel voices". In: *PloS one* 9.3 (2014), e90779.
- [122] Melissa D McCradden et al. "Ethical concerns around use of artificial intelligence in health care research from the perspective of patients with meningioma, caregivers and health care providers: a qualitative study". In: *Canadian Medical Association Open Access Journal* 8.1 (2020), E90–E95.
- [123] Peter E McKenna et al. "A Meta-analysis of Vulnerability and Trust in Human-Robot Interaction". In: *ACM Transactions on Human-Robot Interaction* (2024).
- [124] Roger GT Mello, Liliam F Oliveira, and Jurandir Nadal. "Digital Butterworth filter for subtracting noise from low magnitude surface electromyogram". In: *Computer methods and programs in biomedicine* 87.1 (2007), pp. 28–35.
- [125] Linda Miller et al. "More Than a Feeling—Interrelation of Trust Layers in Human-Robot Interaction and the Role of User Dispositions and State Anxiety". In: *Frontiers in psychology* 12 (2021), p. 378.
- [126] Hamidreza Modares et al. "Optimized assistive human–robot interaction using reinforcement learning". In: *IEEE transactions on cybernetics* 46.3 (2015), pp. 655–667.
- [127] Jahangir Moini, Anthony LoGalbo, and Raheleh Ahangari. *Foundations of the Mind, Brain, and Behavioral Relationships: Understanding Physiological Psychology*. Elsevier, 2023.
- [128] Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

-
- [129] Saeid Nahavandi. "Trust in autonomous systems-iTrust lab: Future directions for analysis of trust with autonomous systems". In: *IEEE Systems, Man, and Cybernetics Magazine* 5.3 (2019), pp. 52–59.
- [130] Matthew A Napierala. "What is the Bonferroni correction". In: *Aaos now* 6.4 (2012), p. 40.
- [131] Tatsuya T Nomura, Dag Sverre Syrdal, and Kerstin Dautenhahn. "Differences on social acceptance of humanoid robots between Japan and the UK". In: *Procs 4th int symposium on new frontiers in human-robot interaction*. The Society for the Study of Artificial Intelligence and the Simulation of ... 2015.
- [132] Michael Novitzky et al. "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*. 2018, pp. 203–204.
- [133] Jihoon Oh, So-Yeong Jeong, and Jaeseung Jeong. "The timing and temporal patterns of eye blinking are dynamically modulated by attention". In: *Human movement science* 31.6 (2012), pp. 1353–1365.
- [134] Shayegan Omidshafiei et al. "Deep decentralized multi-task multi-agent reinforcement learning under partial observability". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 2681–2690.
- [135] Ugo Pagallo. "Apples, oranges, robots: four misunderstandings in today's debate on the legal status of AI systems". In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2133 (2018), p. 20180168.
- [136] Ugo Pagallo. "From automation to autonomous systems: A legal phenomenology with problems of accountability". In: *26th International Joint Conference on Artificial Intelligence, IJCAI 2017*. International Joint Conferences on Artificial Intelligence. 2017, pp. 17–23.
- [137] Sotirios Panagou, W Patrick Neumann, and Fabio Fruggiero. "A scoping review of human robot interaction research towards Industry 5.0 human-centric workplaces". In: *International Journal of Production Research* 62.3 (2024), pp. 974–990.
- [138] LeeAnn Perkins et al. "Designing for human-centered systems: Situational risk as a factor of trust in automation". In: *Proceedings of the human factors and ergonomics society annual meeting* 54.25 (2010), pp. 2130–2134.

-
- [139] Brianna J. Tomlinson Rachel E. Stuck and Bruce N. Walker. "The importance of incorporating risk into human-automation trust". In: *Theoretical Issues in Ergonomics Science* 23.4 (2022), pp. 500–516. DOI: [10 . 1080 / 1463922X . 2021 . 1975170](https://doi.org/10.1080/1463922X.2021.1975170). URL: [https : / / doi . org / 10 . 1080 / 1463922X . 2021 . 1975170](https://doi.org/10.1080/1463922X.2021.1975170).
- [140] PL Patrick Rau, Ye Li, and Dingjun Li. "Effects of communication style and culture on ability to accept recommendations from robots". In: *Computers in Human Behavior* 25.2 (2009), pp. 587–595.
- [141] Yosef S Razin and Karen M Feigh. "Committing to interdependence: Implications from game theory for human–robot trust". In: *Paladyn, Journal of Behavioral Robotics* 12.1 (2021), pp. 481–502.
- [142] GII Research. *Middle East & Africa Industrial Robotics Market Analysis*. Accessed: 2 March 2025. 2024. URL: [https : / / www . giiresearch . com / report / blw1565996 - middle - east - africa - industrial - robotics - market - by . html](https://www.giiresearch.com/report/blw1565996-middle-east-africa-industrial-robotics-market-by.html).
- [143] Grand View Research. *Middle East and Africa Industrial Robotics Market Outlook*. Accessed: 2 March 2025. 2024. URL: [https : / / www . grandviewresearch . com / horizon / outlook / industrial - robotics - market / mea](https://www.grandviewresearch.com/horizon/outlook/industrial-robotics-market/mea).
- [144] René Riedl and Andrija Javor. "The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging." In: *Journal of Neuroscience, Psychology, and Economics* 5.2 (2012), p. 63.
- [145] Paul Robinette et al. "Overtrust of robots in emergency evacuation scenarios". In: *2016 11th ACM/IEEE international conference on human-robot interaction (HRI)*. IEEE. 2016, pp. 101–108.
- [146] International Federation of Robotics. *World Robotics Report 2022: Industrial Robots*. Accessed: 2 March 2025. 2022. URL: [https : / / ifr . org / downloads / press2018 / 2022 _ WR _ extended _ version . pdf](https://ifr.org/downloads/press2018/2022_WR_extended_version.pdf).
- [147] Julian Rode. "Truth and trust in communication: Experiments on the effect of a competitive context". In: *Games and Economic Behavior* 68.1 (2010), pp. 325–338.
- [148] Lucero Rodriguez Rodriguez et al. "A review of mathematical models of human trust in automation". In: *Frontiers in Neuroergonomics* 4 (2023), p. 1171403.

-
- [149] Alessandra Rossi et al. "Evaluating people's perceptions of trust in a robot in a repeated interactions study". In: *International Conference on Social Robotics*. Springer. 2020, pp. 453–465.
 - [150] Julian B Rotter. "Generalized expectancies for interpersonal trust." In: *American psychologist* 26.5 (1971), p. 443.
 - [151] Ericka Rovira et al. "Displaying contextual information reduces the costs of imperfect decision automation in rapid retasking of ISR assets". In: *Human factors* 56.6 (2014), pp. 1036–1049.
 - [152] Hamed Saeidi and Y Wang. "Trust and self-confidence based autonomy allocation for robotic systems". In: *2015 54th IEEE Conference on Decision and Control (CDC)*. IEEE. 2015, pp. 6052–6057.
 - [153] Maha Salem et al. "Would you trust a (faulty) robot? Effects of error, task type and personality on human-robot cooperation and trust". In: *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2015, pp. 1–8.
 - [154] Julian Sanchez et al. "Understanding reliance on automation: effects of error type, error distribution, age and experience". In: *Theoretical issues in ergonomics science* 15.2 (2014), pp. 134–160.
 - [155] Nathan E Sanders and Chang S Nam. "Applied quantitative models of trust in human-robot interaction". In: *Trust in Human-Robot Interaction*. Elsevier, 2021, pp. 449–476.
 - [156] Tracy Sanders et al. "A model of human-robot trust: Theoretical model development". In: *Proceedings of the human factors and ergonomics society annual meeting*. Vol. 55. 1. SAGE Publications Sage CA: Los Angeles, CA. 2011, pp. 1432–1436.
 - [157] Tracy Sanders et al. "The relationship between trust and use choice in human-robot interaction". In: *Human factors* 61.4 (2019), pp. 614–626.
 - [158] Kristin Schaefer. "The perception and measurement of human-robot trust". In: (2013).
 - [159] Kristin E Schaefer. "Measuring trust in human robot interactions: Development of the "trust perception scale-HRI"". In: *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016, pp. 191–218.
 - [160] Kristin E Schaefer et al. "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems". In: *Human factors* 58.3 (2016), pp. 377–400.

-
- [161] Annett Schirmer et al. "Angry, old, male—and trustworthy? How expressive and person voice characteristics shape listener trust". In: *Plos one* 15.5 (2020), e0232431.
- [162] Claudia R. Schneider et al. "The effects of communicating scientific uncertainty on trust and decision making in a public health context". In: *Judgment and Decision Making* 17.4 (2022), 849–882. DOI: [10 . 1017 / S1930297500008962](https://doi.org/10.1017/S1930297500008962).
- [163] Isabel Schwaninger, Geraldine Fitzpatrick, and Astrid Weiss. "Exploring trust in human-agent collaboration". In: *Proceedings of 17th European Conference on Computer-Supported Cooperative Work*. European Society for Socially Embedded Technologies (EUSSET). 2019.
- [164] Sarah Strohkorb Sebo, Priyanka Krishnamurthi, and Brian Scassellati. "'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. 2019, pp. 57–65.
- [165] Ammar Ahmed Siddiqui et al. "Burden of cancer in the Arab world". In: *Handbook of healthcare in the Arab world* (2021), pp. 495–519.
- [166] Michael L Slepian and Evan W Carr. "Facial expressions of authenticity: Emotion variability increases judgments of trustworthiness and leadership". In: *Cognition* 183 (2019), pp. 82–98.
- [167] Harold Soh et al. "Multi-task trust transfer for human–robot interaction". In: *The International Journal of Robotics Research* 39.2-3 (2020), pp. 233–249.
- [168] Harold Soh et al. "The Transfer of Human Trust in Robot Capabilities across Tasks." In: *Robotics: Science and Systems*. 2018.
- [169] Charlene K Stokes et al. "Accounting for the human in cyberspace: Effects of mood on trust in automation". In: *2010 International Symposium on Collaborative Technologies and Systems*. IEEE. 2010, pp. 180–187.
- [170] Rachel E. Stuck, Brittany E. Holthausen, and Bruce N. Walker. "Chapter 8 - The role of risk in human-robot trust". In: *Trust in Human-Robot Interaction*. Ed. by Chang S. Nam and Joseph B. Lyons. Academic Press, 2021, pp. 179–194. ISBN: 978-0-12-819472-0. DOI: [https : / / doi . org / 10 . 1016 / B978 - 0 - 12 - 819472 - 0 . 00008 - 3](https://doi.org/10.1016/B978-0-12-819472-0.00008-3). URL: [https : / / www . sciencedirect . com / science / article / pii / B9780128194720000083](https://www.sciencedirect.com/science/article/pii/B9780128194720000083).

-
- [171] Rachel E Stuck, Brittany E Holthausen, and Bruce N Walker. "The role of risk in human-robot trust". In: *Trust in human-robot interaction*. Elsevier, 2021, pp. 179–194.
- [172] Richard S Sutton. "Reinforcement learning: An introduction". In: *A Bradford Book* (2018).
- [173] Dag Sverre Syrdal, Tatsuya Nomura, and Kerstin Dautenhahn. "The Frankenstein Syndrome Questionnaire—Results from a quantitative cross-cultural survey". In: *International conference on social robotics*. Springer. 2013, pp. 270–279.
- [174] Tetsuya Tanioka et al. "Intentional observational clinical research design: Innovative design for complex clinical research using advanced technology". In: *International Journal of Environmental Research and Public Health* 18.21 (2021), p. 11184.
- [175] Ilaria Torre and Laurence White. "Trust in vocal human–robot interaction: Implications for robot voice design". In: *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers* (2021), pp. 299–316.
- [176] Ilaria Torre, Laurence White, and Jeremy Goslin. "Behavioural mediation of prosodic cues to implicit judgements of trustworthiness". In: *Speech Prosody 2016*. ISCA. 2016.
- [177] Naveen Vemuri and Naresh Thaneeru. "Enhancing Human-Robot Collaboration in Industry 4.0 with AI-driven HRI". In: *Power System Technology* 47.4 (2023), pp. 341–358.
- [178] Lin Wang et al. "When in Rome: the role of culture & context in adherence to robot recommendations". In: *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2010, pp. 359–366.
- [179] Suzhen Wang et al. "Research on optimization of random forest algorithm based on spark". In: *Computers, Materials & Continua* 71.2 (2022), pp. 3721–3731.
- [180] Auriel Washburn et al. "Robot errors in proximate HRI: how functionality framing affects perceived reliability and trust". In: *ACM Transactions on Human-Robot Interaction (THRI)* 9.3 (2020), pp. 1–21.
- [181] Induni N Weeraratna, David Raymond, and Anurag Luharia. "Human-Robot Collaboration for Healthcare: A Narrative Review". In: *Cureus* 15.11 (2023).

-
- [182] Alex Wong, Anqi Xu, and Gregory Dudek. "Investigating Trust Factors in Human-Robot Shared Control: Implicit Gender Bias Around Robot Voice". In: *2019 16th Conference on Computer and Robot Vision (CRV)*. IEEE. 2019, pp. 195–200.
 - [183] Anxing Xiao et al. "Robotic guide dog: Leading a human with leash-guided hybrid physical interaction". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2021, pp. 11470–11476.
 - [184] Yaqi Xie et al. "Robot capability and intention in trust-based decisions across tasks". In: *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2019, pp. 39–47.
 - [185] Xprize. *Ana Avatar Xprize Semifinals Selection Video: Dr. Trina*. 2021. URL: <https://www.youtube.com/watch?v=G2yamXSizDQ&t=43s>.
 - [186] Anqi Xu and Gregory Dudek. "Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations". In: *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE. 2015, pp. 221–228.
 - [187] Caiyue Xu et al. "Trust Recognition in Human-Robot Cooperation Using EEG". In: *arXiv preprint arXiv:2403.05225* (2024).
 - [188] Jin Xu and Ayanna Howard. "How much do you trust your self-driving car? exploring human-robot trust in high-risk scenarios". In: *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE. 2020, pp. 4273–4280.
 - [189] Rosemarie E Yagoda and Douglas J Gillan. "You want me to trust a ROBOT? The development of a human-robot interaction trust scale". In: *International Journal of Social Robotics* 4 (2012), pp. 235–248.
 - [190] Yuchen Yan, Haotian Su, and Yunyi Jia. "Measuring Human Comfort in Human-Robot Collaboration via Wearable Sensing". In: *IEEE Transactions on Cognitive and Developmental Systems* (2024).
 - [191] Holly A Yanco, Jill L Drury, and Jean Scholtz. "Beyond usability evaluation: Analysis of human-robot interaction at a major robotics competition". In: *Human-Computer Interaction* 19.1-2 (2004), pp. 117–149.
 - [192] Holly A Yanco et al. "Methods for developing trust models for intelligent systems". In: *Robust intelligence and trust in autonomous systems* (2016), pp. 219–254.

-
- [193] Boling Yang et al. "Competitive physical human-robot game play". In: *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. 2021, pp. 242–246.
 - [194] Chuang Yu et al. "Robot theory of mind with reverse psychology". In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. 2023, pp. 545–547.
 - [195] Zahra Zahedi et al. "Trust-aware planning: Modeling trust evolution in longitudinal human-robot interaction". In: *arXiv preprint arXiv:2105.01220* (2021).
 - [196] Rong Zhang et al. "A reinforcement learning method for human-robot collaboration in assembly tasks". In: *Robotics and Computer-Integrated Manufacturing* 73 (2022), p. 102227.
 - [197] Tengting Zhang and Hongwei Mo. "Reinforcement learning for robot research: A comprehensive review and open issues". In: *International Journal of Advanced Robotic Systems* 18.3 (2021), p. 17298814211007305.
 - [198] Wenxi Zhang, Willow Wong, and Mark Findlay. "Trust and robotics: a multi-staged decision-making approach to robots in community". In: *AI & SOCIETY* (2023), pp. 1–16.
 - [199] Yinsu Zhang et al. "In Gaze We Trust: Comparing Eye Tracking, Self-report, and Physiological Indicators of Dynamic Trust during HRI". In: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*. 2024, pp. 1188–1193.
 - [200] Yi Zhu et al. "Complexity-Driven Trust Dynamics in Human-Robot Interactions: Insights from AI-Enhanced Collaborative Engagements". In: *Applied Sciences* 13.24 (2023), p. 12989.
 - [201] Sebastian Zörner et al. "An immersive investment game to study human-robot trust". In: *Frontiers in Robotics and AI* 8 (2021), p. 644529.